普通高等教育“十三五”规划教材

统计学原理

马立平　张玉春　主编

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书是作者在多年统计学教学工作经验的基础上编写出来的，用通俗易懂、深入浅出的语言阐述统计学的基本思想与方法，具有较强的实用性和可操作性。读者通过学习可以掌握统计学中基本的、常用的统计方法的原理与思想，并能够在此基础上正确地使用统计方法进行统计分析与研究。全书共 16 章，分别为导论、数据收集的方法、数据的预处理与分组整理、数据特征的统计量描述、数据资料的图形显示、统计指标与多指标综合评价、概率抽样方法与抽样分布、参数估计、参数的假设检验、方差分析、列联分析与对应分析、相关与回归分析、时间数列的描述性分析、时间数列的构成与预测、聚类分析和判别分析、主成分与因子分析。

本书可作为高等学校统计学基础课程的教材，也可作为感兴趣的读者学习统计学入门知识的参考书。

图书在版编目(CIP)数据

统计学原理 / 马立平, 张玉春主编. —北京：电子工业出版社，2018.11
ISBN 978-7-121-34216-5

I．①统… II．①马… ②张… III．①统计学－高等学校－教材 IV．①C8

中国版本图书馆 CIP 数据核字(2018)第 099218 号

策划编辑：王二华
责任编辑：王二华　　文字编辑：蔡馥羽
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1092　1/16　印张：17.75　　字数：450 千字
版　　次：2018 年 11 月第 1 版
印　　次：2018 年 11 月第 1 次印刷
定　　价：46.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010)88254532。

前　言

统计学是收集、分析、表述和解释数据的方法论科学。通过数据资料对现象进行数量方面的分析，不仅能够更客观地认识研究对象，也能使对问题的认识与研究更加深入，统计方法的应用在某种程度上影响着实际工作与理论研究的水平。统计学方法可以应用到几乎各个领域中，其既可用于对社会现象数量方面的研究，也可用于对自然现象数量方面的研究，在科学研究与社会经济管理中发挥了重要的作用。相信读者在学习统计学并实际应用之后，能够更深刻地感受到这一点。

我们编写本书的目的是，让读者可以通过学习掌握统计学中最基本的、常用的统计方法的基本原理与思想，并能够在此基础上正确地使用统计方法进行统计分析与研究。本书在撰写时特别注意突出以下几方面的特点。

第一，为使初学统计学的读者尽快了解统计研究的思维方式、掌握统计分析的基本方法并进行应用，本书编写时着重用通俗易懂、深入浅出的语言阐述统计学的基本思想与方法，力求强调统计思想，略掉了一些数学证明和公式推导，在对各种具体方法做必要的阐述之后，配备有实例说明其基本思想与方法应用，并用图表等比较直观的形式进行解释。

第二，突出实际问题分析与统计研究中的工作顺序与知识逻辑关系。本书在结构上共分为 5 个部分，即统计数据及特征描述、单一指标的统计推断、变量之间关系的统计推断、时间数列分析与预测、统计聚类与数据降维。在内容上不仅包括描述统计，也包括推断统计；从分析方法方面不仅包括传统意义上的统计学原理的基本内容，也包括实际分析研究中常用的多元统计分析的方法与应用, 力求使读者学习后能够系统地、较为完整地掌握统计方法论的总体框架，从而在学习了解统计学的基本理论、思想与方法的同时，能够把握何时、用何种方法、对何种问题进行统计分析与研究。

第三，统计学是通过对大量数据的分析与研究发现问题、得出结论的方法论科学，具有很强的应用价值。为了避免读者学完之后只掌握了统计学基本思想与方法而无法实际操作，本书编写时力求具有较强的实用性和可操作性，在介绍方法的基础上，结合统计软件全面、系统地介绍统计计算与分析过程及其技术实现，针对具体问题进行分析并对计算结果进行解读，以提高读者统计分析方法的实际操作能力与水平。

本书由马立平、张玉春编写。本书的编写是建立在作者多年本科及研究生统计学教学经验与体会的基础上的，同时也参考了许多统计学专著和教材，得到了电子工业出版社王二华老师的大力支持和帮助，受到了北京市高层次创新创业人才支持计划教学名师项目的资助，在此一并表示感谢。尽管编写此书时我们投入了许多时间和精力，但书中难免还会存在一些不尽人意之处，真诚欢迎广大读者提出宝贵意见，在此先行表示感谢。

编　者

目　录

第五部分　统计聚类与数据降维

第一章　导　　论

第一节　统计学的产生与发展

人类的统计活动最早可追溯到远古的原始社会，历史源远流长。可以说，自从人类有了数的概念，有了计数的需求，也就有了统计活动。但是，将统计活动、统计实践上升到理论并形成“统计学”这门科学却是近代的事。虽然对于统计学产生于什么年代这个问题人们的看法不尽一致，但多数人认为，统计学兴起于 17 世纪，距今有 300 多年的历史。

一、统计学的产生

17 世纪中叶，西方社会出现了人们有意识地、比较系统地用数字语言表述问题，并从数量的角度探索客观事物变化规律的研究活动。当时最著名的、最具有代表性的是政治算术学派和国势学派。

政治算数学派的代表人物是英国的学者威廉·配第(W. Petty)。17 世纪 70 年代，威廉·配第的著作《政治算术》问世，书中威廉·配第以劳动价值论理论为基础，对当时的英国、荷兰、法国之间的国情、国力(主要是经济实力)进行了数量上的对比分析，并以此为依据，为当时英国的社会经济的发展出谋划策。这是历史上首次明确地从数量的角度、用大量的数据资料分析问题，揭示了一些经济学的科学原理，研究了许多经济范畴中的经济关系。无论是在古典政治经济学学科，还是在统计学学科的发展史上，《政治算术》都可以称得上一部具有奠基意义的重要著作，马克思曾称威廉·配第是政治经济学之父，是统计学的创始人。在关于统计史的研究中，人们一般把以威廉·配第为代表的关于社会经济现象的“算术”式的研究，称为“政治算术”。从统计学科看《政治算术》的意义主要表现在人们所研究、关注的问题这方面。威廉·配第在其著作的序言里写道：我进行这种工作所使用的方法在目前还不是常见的，因为我不采用比较级或最高级的词语进行思辨式的议论，相反地，采用用数字、重量和尺度来表达自己想说的问题的方法。自威廉·配第后的 200 年内，以用数量方法研究社会经济问题为基本特征的政治算术模式，是统计学发展的主流。

除威廉·配第外，政治算术还有一位重要的人物约翰·格朗特(J. Graunt)，他的主要工作是对伦敦市 50 多年的人口出生和死亡数进行计算分析和研究。1662 年约翰·格朗特写出了代表性的著作《关于死亡表的自然观察与政治观察》，该著作通过对人口变动数据的分析，揭示了一系列的人口变化规律。从此，统计的含义从由记述为主，转变为从量的方面说明并分析国家的重要事项，为统计学作为从数量方面认识事物的一种方法开辟了广阔的研究与应用前景。政治算术学派第一次有意识地运用可度量的方法，力求把自己的论证建立在具体的数字基础上，其在统计发展史上具有重要的地位。但是，其毕竟还处于统计核算的初始阶段，从现在的视角看，它只是用最简单的算术方法对社会经济现象进行计量和比较。

统计学的另一个重要起源是概率论。概率论虽然起源于赌博游戏，但真正意义上的概率

论，是从 17 世纪开始的。在早期从事概率论研究的众多学者中，拉普拉斯是古典概率的集大成者，他给出了概率数学的古典解释，建立了严密的概率数学体系。与“政治算术”研究的是简单的、确定的数量关系不同，概率统计研究的则是复杂的、随机性的现象。概率论的出现极大地充实和深化了数量问题研究的内容，以概率论为基础的统计学也很快进入了一个新的发展时期。

二、统计学的发展

凯特勒(A. Quetelet)是统计学发展史上承前启后的重要人物，人们称他为“近代统计学之父”。他在统计学上做出的突出贡献是把概率论全面引进“政治算术”，引入到各种社会经济问题的研究当中，大大推动了概率论和数学方法的应用，促进了数量研究由“算术”水平向“数理”阶段的迅速转化。

自文艺复兴以后，人们已经注意到，当诸如玩纸牌、掷骰子等赌博活动大量进行之后，会出现某种类型的规律性，而概率论最早就是研究这种规律性的产物。当然，概率论的产生和形成在 16～18 世纪，当时与统计学关联性不强，统计学也很少将概率论应用到自己的领域，真正将统计学与概率论结合起来的是凯特勒。凯特勒在自己的研究工作中，首次在社会科学的范畴提出了大数律思想，把统计学的理论建立在大数律的基础上，认为一切社会现象都会受到大数律的支配，他不仅把概率统计的方法引入人口、领土、政治、农业、工业、商业、道德等社会领域，还把概率统计的方法引入天文、气象、地理、动物、植物等自然领域。他的关于概率统计的方法是可以应用于任何事物数量研究的最一般方法的思想，对统计学的发展具有重大意义。

19 世纪后半期，统计学在生物遗传学、农业田间试验等领域都取得了创新性的成果。例如，生物统计学的主创者高尔顿(F. Galton)利用正态法则研究优生学、遗传学等问题，先后提出了“百分位数”、“中位数”、“四分位数差”、相关与回归等概念及计算方法。而皮尔逊(K. Pearson)则系统地发展了高尔顿的相关与回归理论，研究复相关和偏相关，研究出极大似然估计方法，导出了卡方分布。以皮尔逊为代表的统计学家，通过大量观察和以正态分布为基础的关于总体分布曲线的研究，确立了大样本统计理论，从而奠定了描述统计学的框架体系。

进入 20 世纪后，随着新的统计思想和统计方法的大量涌现，带有归纳性质的统计推断逐渐占据了统计学的主流地位。从苏歇米尔斯(J. Sussmilch)提出大数法则开始到 20 世纪初的这段时期，大量观察法一直是统计的核心思想，直到 1908 年戈塞特(W. Gosset)导出了重要的 t 分布，统计学逐渐实现了由描述统计阶段发展到推断统计阶段，实现了由大样本统计向小样本统计理论的转变。费雪(R. Fisher)开辟了方差分析、试验设计等统计分枝，论证了相关系数的抽样分布，提出了 t 检验、F 检验、相关系数检验等理论与方法，因而在统计学发展史上有着很高的地位。此后，内曼(J. Neyman)和皮尔逊共同完善了现代统计学的核心内容，即区间估计和假设检验的理论，瓦尔德(A.Wald)提出的统计决策理论和质量检验的“序贯分析”，进一步推动了统计学研究和应用的范围。到 20 世纪五六十年代后，随着统计学的发展，稳健统计、时间数列、抽样理论、统计诊断、探索性分析、贝叶斯统计等取得了重要的进展。随着网络信息技术的发展，面对大数据时代，统计学与统计方法将面临着又一次的革命和飞跃。

20 世纪以来，统计学的发展表现出三个明显的趋势：(1)随着数学的发展，统计学依赖和吸收数学方法的程度越来越深入，发展越来越迅速；(2)统计学方法应用领域越来越广泛，向其

他学科领域的渗透越来越深入，以统计学为基础的边缘学科不断形成；(3)随着应用的日益广泛和深入，特别是计算机的发展，统计学发挥了重要的作用，而且将会发挥越来越大的作用。

第二节 统计学的性质与统计方法

一、什么是统计学

提到“统计”这个词，人们自然会将其和统计数字、统计数据联想到一起，如人口总数、国内生产总值、物价指数、收入水平、消费支出等。从统计学的产生和发展可以看到，统计学是随着人类社会的发展和社会管理的需要而不断发展起来的，它是一门以大量现象的数量方面为研究对象的认识方法论科学。

随着统计方法在各个领域的应用，统计学已发展成为具有多个分支学科的大家族，对统计学的认识由于应用和研究的角度不同而不尽相同。正因为如此，即使是统计学家，也对统计学的定义给出了不同的回答，有着不同的表述。在这之中，具有代表性的、被普遍接受的是不列颠百科全书的定义：“统计学是收集、分析、表述和解释数据的科学。”统计学的定义告诉我们，统计学是关于数据的方法论科学，这意味着统计学离不开数据，其可以表现在统计研究过程的各个环节中。

从统计研究的整个过程来看，进行统计研究首先需要搜集能够反映或说明客观现象特征的数据资料，这是统计活动的首要的、也是最基本的环节。没有统计数据，统计方法就失去了用武之地，而如何取得可靠的、高质量的统计数据则是统计学研究的重要内容之一。

有了统计数据后，如何进行数据的分析则是统计学的核心内容。统计分析是对已有的数据资料，通过科学的统计方法探索数据的内在规律、提取有价值信息的过程，其目的是形成一个对研究对象具有概括性的、全面的、整体的数量描述。分析数据所用的方法可分为描述性统计方法和统计推断方法，其中描述性统计是对数据的分布形态、数量特征和随机变量之间的关系等进行估计和描述的方法，主要包括对数据的集中趋势、离中趋势和变量间相关关系等内容的概括性描述方法。而推断统计主要研究如何根据样本数据去推断总体数量特征，它是在对样本数据进行描述的基础上，对统计总体的未知数量特征做出以概率形式表述的推断。当然，分析数据之前，需要对统计数据进行加工和整理，即数据整理，目的是使统计数据系统化、条理化，以符合统计分析的要求并得到正确的分析结果。统计整理在统计活动中处于承上启下的位置，一方面它是统计搜集资料工作的延续，另一方面又是统计分析的前提，是统计工作的必要环节。数据整理的内容包括：对所搜集的数据进行甄别与筛选、对数据进行审核与修正、对数据进行分组分类、编制数据频数分布表、计算统计指标、数据的图示表现与可视化等，通过数据整理可以帮助分析研究人员发现数据的初步特征，或者方便他人看懂数据所要表达的问题。

数据的解释是对分析的结果进行说明和进一步的分析，阐明分析结果所隐含的事物的特征，从数据中或分析结果里得出关于研究对象发展变化规律性的结论等。

二、统计学的性质与研究对象

统计学的发展历史是从社会经济现象的数量开始的，随着统计方法的不断完善和应用领域的不断拓展，统计学得到了快速的发展。

总体来看，统计学的性质主要表现在以下几个方面。

第一，统计学是一门关于数据的科学，它对客观事物的研究是通过对客观事物数量方面的研究来进行的，包括研究对象的数量状态、数量关系和数量变化规律等，可以说统计学的研究对象是数据。无论是社会科学还是自然科学，只要出现大量数据的地方，就需要统计学。从认识论的角度，任何事物都是数量和质量的统一体，如果数据资料真实、准确、可靠，且统计分析方法运用得当，那么通过客观事物的数量方面特征就可以正确地认识客观事物的特征与其发展变化的规律。

第二，统计学的研究对象是总体，而不是个体；是客观事物的规律，而不是偶然出现的某一现象。统计学是对大量同类现象的数量方面进行描述并对总体进行推断，对单个个体现象的分析与研究不是统计的研究对象。只有通过对大量的现象，或对某一现象进行多次重复的观察，才有可能找到统计关系和统计规律。按认识论由个体数量到总体数量的认识逻辑，统计研究的对象是总体的数量方面，但需要从个体的数量方面入手。

第三，统计学的研究对象是不确定现象，表现为随机变量。不确定现象的存在是因为在一些偶然的、随机因素的影响下，客观事物的实际数量表现存在一定程度的不可确知性。现实中，太多的现象都是不确定的现象，如总体上良好的生活习惯可以延长我们的寿命，但具体到某一个具体的人来说，其寿命则是很难预先确定的，可能会出现一个生活习惯不好经常吸烟、喝酒并且不锻炼的人比一个生活习惯良好的人活得还长的现象，当然也有可能短。可以说，人的寿命就是一个随机变量，它除了受到遗传基因的影响外，还会受到生活习惯、生活质量等很多随机因素的影响。但同时我们也知道，虽然一个人的寿命有一定的随机性，但从总体上看，我国公民的预期寿命是非常稳定的，且女性的预期寿命高于男性的预期寿命，这就是在随机性之中的规律性。

第四，统计学的基本方法是归纳推断。统计对总体的认识有两种途径，一是在掌握关于研究对象的全部数据资料的基础上，运用算术方法和统计描述手段就可达到认识总体的目的；另一个是随机抽取样本，并依据样本的数据，利用归纳推断的方法，对总体特征进行推断。从应用的经济性、时效性、实用性和可行性等方面考虑，利用样本数据进行总体推断的优势比较明显。

三、统计的基本方法

统计方法是指统计学研究和认识客观事物总体数量方面的各种方法。从研究主体或工作过程的角度来说，统计学研究要经历资料的搜集、整理和分析的工作过程，或称为统计调查工作、统计整理工作和统计分析工作。在这一过程中的各个阶段，都会有专门的统计学方法。从总体上说，统计学的基本方法有大量观察法、综合指标法和归纳推断法。

(一)大量观察法

大量观察法是统计学研究的基本方法，它是一种从总体出发对研究对象的全部数据或足够多的数据进行观察和分析研究的方法。

大量观察法之所以是统计学研究的基本方法，这是由统计学的研究对象及研究目的所决定的。统计学的研究对象是总体的数量方面，是由大量数据构成的。研究对象的数量方面会受到诸多因素的影响，我们可以将这些因素分为性质不同的两大类：一类源于研究对象的基本性质及一般条件的共同性因素，这类因素对所有个体单位都会发生作用，是研究对象总体

数量规律性存在的基础。另一类源于研究对象的次要性质或偶然因素、随机因素在个体单位上发生的作用，正是由于这类因素的作用，使得各个单位在数量表现上存在差异，各不相同，或多或少地掩盖了研究对象的规律性。大量观察法的意义就在于，在全部或足够多的数据的基础上，去除掉偶然或随机因素的作用，突出共同性因素的作用，从而显示出总体相对稳定的数量特征和数量关系，即数量规律性。

(二)综合指标法

综合指标法是统计学研究中直接表现研究对象总体数量特征的最基本的统计方法、统计手段或工具。综合指标按其一般表现形式可分为总量指标、相对指标和平均指标三大类。其中总量指标的基本来源是对原始数据的整理汇总，以其为基础，可利用多种方法计算出各种派生的相对指标和平均指标。

在统计学的研究中，综合指标有着重要的意义。它不仅可以概括地表明研究对象的规模、总量，也可以表现研究对象的一般水平、内在结构和比例特征，它是对研究对象数量方面的一种测度。在指标的基础上，我们还可以进行统计指标及指标关系的分析，也就是对数量特征、数量关系、数量界限及数量规律性的分析。可以说综合指标在统计分析与研究中承担着重要的作用。

(三)归纳推断法

由于种种主、客观方面的原因，我们经常会碰到所研究的对象的范围大于实际可能掌握的范围的情况，这时要认识总体的数量特征，就需要应用统计推断法。统计推断法是指以一定的置信水平，根据样本数据来估计总体数量特征的归纳推断方法，它是现代统计的重要方法。统计推断的一个重要特点是它不能对问题做出绝对肯定的结论，只能在一定的置信水平下，做出能满足研究精度的弹性结论。

四、统计学科体系

目前统计学已经形成了由理论统计学、应用统计学、统计学史等若干分支组成的完整的学科体系。

(一)理论统计学

理论统计学是侧重于研究统计学的方法论和基础理论，以解决统计学学科发展中的重大问题为目标的统计学分支。其最基本内容包括以下几个方面。

(1)统计估计。统计估计是统计学的核心内容之一。它包括两个方面的内容，一是在总体分布已知时，对总体未知参数或参数组合的函数进行估计；二是在总体分布未知时，对有关分布的特征数字及分布密度进行估计；其研究的重点内容包括估计量的确定和对估计量的评价等。

(2)假设检验。假设检验是统计学的另一重要内容。它是指根据样本资料，对总体参数的某种假设命题进行检验和推断。其研究的重点在于检验统计量的构造、假设检验的原理和检验效率等问题。

(3)抽样调查。抽样调查是搜集统计资料的基本手段与方法之一，也是统计学的一个重要分支。其研究的重点在于抽样方案的设计、样本的抽样方法、抽样分布、抽样效果与抽样误差等问题。

(4)试验设计。试验设计主要研究如何安排试验方案、获取试验数据，以及如何对试验结果进行分析等问题。

(5)非参数统计。非参数统计主要研究总体分布未知或不依赖于总体分布及非总体参数的各种统计问题。

(6)时间数列。时间数列是指按时间顺序排列的一组数据，时间数列方面的主要研究内容包括有时间数列的基本结构、时间数列的分解、自回归过程与参数估计、非线性系统模型和空间序列分析等。

(7)统计决策。统计分析与研究的最终目的是在认识客观规律的基础上作出科学的决策，统计决策部分中的主要内容包括风险函数、损失函数、决策标准和决策函数等。

(8)序贯分析。序贯分析是指在得出分析结论之前，视具体的观察结果决定决策方案的选择，其包括抽样方案、序列检验统计量、判别风险等内容。

(9)多元统计。多元统计是针对多维随机变量的统计分析方法，是从经典统计学中发展起来的一个分支。多元统计是当总体的分布是多维(多元)概率分布时，处理该总体的理论和方法，它能够在多个对象和多个指标互相关联的情况下分析它们的统计规律，是一种综合分析方法。多元统计主要包括多元正态分布及其抽样分布、多元正态总体的均值向量和协方差阵的假设检验、多元方差分析、判别分析、典型相关分析、主成分分析、因子分析、聚类分析、多元回归等内容。

(10)统计诊断。统计诊断是近几十年来发展起来的一个统计学的新的领域，统计诊断主要研究观察数据、统计模型、统计推断方法的合理性问题，并对诊断中发现的缺陷进行治理和改进。

(11)稳健统计。稳健统计主要研究当理论分布与实际分布不一致时，如何确定不敏感的统计方法等。

(12)探索性分析。探索性分析就是通过对观察数据进行详细的考察，力求挖掘出数据本身的结构和特征，然后在此基础上建立分析模型。

(二)应用统计学

应用统计学就是运用统计思想和方法，处理实践中遇到的各种问题的理论与方法。应用统计学大体可以分为以下几类。

(1)统计计算方法。它把统计方法、数学计算方法和计算机应用结合起来，重点解决数据处理过程中所碰到的各类计算问题。

(2)应用统计学理论基础。应用统计学理论基础是站在理论统计学角度上的应用统计学，同理论统计学相比，应用统计学带有较强的应用背景。

(3)统计学应用。从实际问题的背景出发，与具体研究对象所在领域学科紧密结合，重点在于如何应用统计方法于实践之中。应用统计学按应用的学科性质不同，可区分为理工科的应用统计学(如统计力学、生物统计学、医学统计学、气象统计学、地理统计学等)和社会科学类的应用统计学(如人口统计学、经济统计学、管理统计学、教育统计学、社会统计学等)。

(4)统计学与其他应用数学学科的结合，形成了新的应用数学方法的基础学科，如博弈论、多目标决策、随机规划等。

第三节 统计的应用领域

目前，统计学的理论与方法已被广泛应用到自然科学和社会科学的众多领域，如在工农业生产和商业活动方面，在社会学和政治学方面，在史学和考古学方面，在物理、化学和生物学方面，在天文地理方面，在交通运输和能源供应方面，在医学和保健方面，在教育和文化方面，在保险和社会福利方面，在自然科学和实验方面等，基本上都要用到统计工具。统计学的理论、方法与相关学科的结合，逐步形成了相关学科的统计学分支，统计学也已经发展成为由若干分支学科组成的学科体系，如经济统计学、生物统计学、物理统计等。可以说，当今年代，无论我们从事什么活动，大多都离不开数据，离不开统计学。即使对于一些复杂现象，我们一时可能难以掌握其变化规律，统计方法也未必是认识和处理问题的唯一途径，但它却可以帮助我们发现随机现象中的必然性，探究隐藏在表面现象背后的规律性，或许也可以给我们提供一些线索，引导我们把研究、分析活动深入下去。

在统计学的应用实践中，经济、管理等领域是统计方法得到较早应用和较多应用的一个领域。经济与管理统计应用中不仅包括宏观领域的经济统计分析，也包括微观领域中的企业商务管理统计的应用等。

在全球商务和经济环境中，有大量的统计信息是可利用的。成功的管理人员和决策制订者一定是那些能够理解信息并有效利用信息的人。而管理统计学，正是运用统计方法，分析和解决企业经营和管理活动中遇到的各种需要解决的问题，帮助做出正确的选择与决策。一般，管理统计中研究的问题主要有市场统计分析、产品试验设计、人员调度、成本预算管理、库存管理、生产控制与管理、风险防范、财务管理等。

统计方法在会计工作中也得到了较好的应用。如会计师事务所想确定列在客户资产负债表上的应收账款是否真正地反映了应收账款的实际金额，就可以用到统计方法。一般应收账款的数量是庞大的，以至于查看和验证每一账户要花费太多的时间和费用。这时，就需要审计人员从账户中选择部分账户作为样本，在对样本账户审查其准确性后，推断并得出有关列在客户资产负债表上的应收账款金额是不是可以接受的结论。在这过程中选取样本、得出结论的过程和方法就需要使用统计学的随机抽样方法和参数估计方法等。

统计方法在理财中的应用。理财顾问是目前一个相对比较新兴的职业，其主要是利用各种各样的统计信息来指导客户、根据客户的实际情况给其相应的投资建议。如在股票投资中，理财顾问要查询阅览各种财务数据，包括股票的价格、盈余比率和股息率，通过对比单个股票的信息和股票市场平均状况信息，得出单个股票的价值是被高估还是被低估，给客户提出买、卖或持有股票的建议等。

统计方法在市场分析中的应用。随着电子技术的发展，超市、专卖店、百货商场等零售企业收款台的电子扫描仪在收款的同时，也起到了收集各种市场调研所需要的各种数据的作用。通过对这些数据的统计分析，可以为商家提供消费者消费行为的基本特征与规律，也可以向制造商出售数据的统计分析摘要，从而提出建议帮助商家和制造商制订关于各种促销的活动计划，诸如特价销售和店内各种商品的陈列方式及位置等。此外，产品经理可以通过对扫描资料和促销活动相关数据资料的统计分析，了解促销活动和销售额之间的关系，掌握促销活动的效果，等等，类似于这种统计分析将有助于各类产品发展战略的制订。

此外，统计方法还可以在产品质量控制中发挥其作用，用于监测生产过程的产出。还有

在对未来的经济状况或对未来经济的某一方面进行预测、判断时，需要使用各种统计信息和统计方法。例如，在预测通货膨胀率时，可以利用有关诸如生产价格指数、失业率、生产能力利用率等统计信息，通过统计模型进行预测。

可以看到，统计学理论与方法是一种重要的、定量分析的方法论与工具，在社会、经济、自然等各领域的研究与实践活动中具有重要的作用，今后也将会发挥出更加重要的作用。但是它不是万能的，不能解决我们想要解决的所有问题，而且如果选用了错误的方法、角度或数据，就可能会得到错误的结论。

能否用统计方法解决具体的问题，首先要看使用统计方法的人是否对所研究的对象有基本的、初步的认识，即定性认识；其次，是否能选择正确的统计方法对可靠的数据进行分析；最后，能否应用研究对象所在学科领域的专业知识对统计分析的结果做出合理的解释和分析。

思考与练习

1．统计学的研究对象是什么？
2．统计学的性质是什么？
3．请举例说明统计方法的作用。
4．统计学的研究方法主要是什么？

第一部分　统计数据及特征描述

统计的研究对象是数据，数据是基础，数据的类型与数据的质量在统计分析中具有相当重要的作用。针对不同类型的数据，我们可以选用不同的分析方法。数据的分布与特征不仅反映了研究对象的基本状况，而且不同的分布特征也决定了不同的分析方法。以下几章将分别介绍数据的收集方法、数据的预处理与分组整理、数据特征的统计量描述、统计图显示、统计指标与多指标评价等方法。

第二章　数据收集的方法

第一节　数据的来源

根据定义，统计学是收集、分析、表述和解释数据的科学，其核心是数据，而这些数据是从哪里来的呢？通常只要肯看，愿意去观察，就或多或少可以得到我们所关注的研究对象的数据。然而得到数据后，我们还要面对数据质量的问题，我们是否应该相信所掌握的数据，并对其进行分析以得出结论呢？好的数据是人们智慧和努力的结果，可以帮助我们客观地认识社会经济与自然现象，做出理性的选择；而坏的数据则会误导我们，甚至使我们误入歧途。面对我们所拥有的数据，应用前，我们首先应该了解：这些数据是从哪里来的？它是否是真实的、可靠的？

从使用者的角度看，统计数据的来源主要有两种渠道：一是来源于直接的调查与实验，即一手数据的来源，也称为直接来源；二是来源于别人调查、实验并经过加工整理后的数据，也称为二手数据或间接来源。从数据本身来源看，其最初的来源都是直接的来源，只不过间接来源是别人直接调查的并经过加工整理之后的数据。

一、数据的间接来源

进行统计调查、做科学实验收集数据，往往需要一定的条件。对于大多数人来讲，我们不太可能或不需要面对所有的问题与需求都采用直接的调查方法去收集第一手的原始数据资料，因而所使用的数据很多都是二手数据，即获取现成的资料。

二手数据很多是公开出版或公开报道的数据，当然，有些是尚未公开出版的数据，一般称这类数据收集的方法为文案调查。

文案调查通常按以下几步进行。

第一，根据研究项目的目的与内容确定所需要资料的类型。例如要进行市场研究，收集反映市场状况的数据，应根据研究目的首先确定收集的是宏观数据还是微观数据，是收集动态的数据资料还是收集静态的数据资料，等等。

第二，寻找资料来源。二手资料的来源很多，最主要的是公开出版的或公开报道的数据，如来自国家和地方统计部门出版的年鉴——《中国统计年鉴》、《中国社会统计年鉴》、《中国工业经济统计年鉴》、《中国农村统计年鉴》、《中国人口统计年鉴》、《中国市场统计年鉴》以及各省、市、地区的统计年鉴等，除此之外，还有行业协会发布的数据和一些权威的研究报告中的数据等。关于世界各国的统计数据有联合国有关部门、机构和各国出版的统计数据等。 当然，除了公开出版的统计数据外，还可以通过其他一些渠道使用一些尚未公开的统计数据，以及广泛分布在各种报纸、杂志、图书、广播、电视传媒、网络、历史文献及著作等中的各种数据资料。

第三，对二手数据资料的追踪查找。在查找所需的资料时，也可以根据与调查研究的项

目有关的著作论文末尾所列的参考文献目录进行追踪查找，还可以利用检索工具(目录、索引和文摘)进行查找。

第四，对数据资料进行加工、整理和补充。二手数据对使用者来说，其优点在于获取资料较为方便、容易，调查费用低。但应注意二手数据资料是其他人或机构为满足其自身的目的或某特殊目的收集的，因此将这些资料用于其他的研究项目时，目的可能不同，往往在时间上、资料的完整性上具有一定的局限性，在数据统计口径与计算方法等方面也许不一定能满足要求，需要进行相应的调整和补充，避免统计数据的误用。

二、数据的直接来源

收集一手数据资料是统计活动的重要内容之一，数据的直接来源主要有两个渠道，一是通过调查或观察，二是基于试验设计基础上的试验。

(一)统计调查

统计调查是取得统计数据的主要来源，也是获得直接统计数据的重要手段。统计实践中最主要的调查方式如下。

1. 全面调查

全面调查是指对研究对象的所有单位均进行调查，搜集研究范围内所有单位数据资料的调查方法。

2. 普查

普查是专门组织的、一次性的全面调查，其主要用来收集某一时点或一定时期内现象总量资料的，利用普查，我们可以获取被研究事物总体的全面情况。从宏观上看，国家通过普查可以摸清一个国家的国情和国力，了解到一个国家人力资源和物质资源的现状及利用情况，如人口普查、经济普查、农业普查等，这对于国家做出重大决策、制定政策和计划及经济与社会发展的长远规划都是不可缺少的。从微观上看，普查也可用于某些小范围的市场调查，如对市场上某种产品的供应、销售及库存的全面调查。

虽然普查需要耗费相对较多的人力、物力和财产，但其作为获取全面资料的有效途径之一，仍然保持着特殊的地位，许多国家都把普查明确规定为一项制度，每隔一定时期就举行一次。普查比任何其他调查方式所收集的资料都更全面、更系统，但国家层面的普查工作牵涉面广、工作量大，需要较多的人力、物力、财力，单独的一个研究机构往往无力承担。从准备调查方案、设计表格、试点、培训普查员，到实施调查以及后续的资料整理和分析，需要较长的时间。另外，大量的社会经济问题研究，并不一定需要通过普查来收集数据资料，如果全都用普查势必得不偿失。

3. 抽样调查

抽样调查是统计实践中应用最广泛的一种调查方式。它是从所研究对象的总体中按照一定的原则抽取出部分单位作为样本进行调查。按照抽选样本的原则，抽样调查可分为主观选择样本的抽样方式和按随机原则抽取样本的概率抽样。

(1)主观抽样调查是指研究人员有意识地按照自己的判断与偏好选择样本进而收集数据的方式，常用的有典型调查、重点调查、配额抽样调查等。

典型调查是根据调查的目的，在对研究对象进行全面分析的基础上，有意识地选出少数具有代表性的或具有典型意义的单位，进行深入细致调查的一种调查方式。典型调查的作用

主要在于：第一，可以对新生事物或典型事例进行深入的分析和研究，查找原因，也可以作为其他统计调查的补充；第二，在一定条件下，可检查普查数据的真实性等。

重点调查是在调查对象中选择那些少数重点单位进行调查，以掌握研究对象的基本情况。所谓重点单位是指在所要关注的数量方面，那些单位数少，但其数值在所有单位中具有举足轻重(即数量占比很大)地位的单位。组织重点调查的关键问题是确定重点单位，重点单位选多选少，要根据调查任务来确定。一般来说，选出的单位尽可能少些，而这些单位又能反映总体的一般情况。这些被选中的单位能提供较为可靠的数据，达到重点调查的目的。

配额调查是市场调查和民意调查实践中通常采用的方法，它是指预先确定各种类型的调查对象在整个样本中所占的比例，然后在这一比例分配的基础上，主观选取具体抽选哪些具体的样本单位。这种方法在抽取样本单位环节时，由于是主观抽样，其抽样简便，费用节约，有很大的使用价值。但该抽样方法很大程度上依赖于调查者的主观判断，因而很难避免由于主观判断而产生的样本的偏差，而且这种偏差是无法客观计算的。

(2)概率抽样调查是按照随机原则在调查对象中抽取样本进而收集数据资料的方式，该方式避免了主观上有意识地抽取样本而导致样本不具有总体代表性的不足，也就是说，被抽中的单位不取决于研究人员的意愿和偏好，而是依赖于客观的机会——概率。当然，每个单位被抽中的概率可以根据抽样的设计事先确定。这种抽样调查方式的优点是，根据概率论的原理可以用一定的置信度控制抽样误差在规定的范围之内，也就是说，用样本推断总体。

(二)试验设计

产品的设计、科技成果的取得，很多需要进行科学试验。科学理论的产生，不仅仅来源于生产实践、社会实践，也来自于科学试验研究。在许多学科中，试验是进行科学研究的重要手段，而科学试验离不开科学的统计试验设计。所谓统计试验设计是为了获取科学试验的各种数据所制订的计划，产生或获取数据后，研究人员可以利用统计方法对其进行分析、探索，发现对试验结果产生影响的因素和影响程度，以便进行进一步的研究。

试验设计一般包括五个环节：首先是进行方案设计，即根据试验的目的，设计试验方案，以保证每一个非被试验因素对被试验对象的作用相同，以突出被试验因素的效应，合理优化试验的次数；其次是试验设计方案的实施；第三是采集数据，即对试验结果进行测量，取得试验数据资料；第四是进行数据分析，即运用统计方法，对试验数据进行分析，形成相应的结论；最后是根据对数据分析所得的结果，得出试验结论。

第二节　统计调查方案的设计

在对原始资料进行搜集即统计调查时，首先需要制订一个科学、合理的调查方案。一般地，一个完整的调查方案需要包括调查目的、调查对象、调查单位、调查项目、调查方法和调查的组织与实施方法等基本内容。

一、调查目的

明确调查目的是设计调查方案的关键所在，只有有了明确的调查目的，才能确定调查的范围、内容和方法，取得高质量的统计数据资料，否则调查时极有可能包括一些无关紧要的或不需要的统计资料，从而降低统计调查的效率，也有可能遗漏一些重要的调查内容，因而

无法满足统计分析与研究的需要，导致调查工作的失败。例如，我国城市住户调查的调查目的是了解城市居民家庭人口、就业、消费、储蓄、手存现金、商品需求及住房变化等情况，为政府研究制定劳动力就业、社会保障与福利、货币流通、商品生产服务的供应等方面的政策提供依据。再例如，普查是一项专门组织的全面调查，组织一项调查需要耗费大量的人力、物力、财力和时间，调查前必须明确调查的目的。我国进行经济普查的目的是全面调查了解我国第二产业和第三产业的发展规模和布局，了解我国产业组织、产业结构、产业技术的现状以及各生产要素的构成，摸清我国各类企业和单位能源消耗的基本情况，建立健全覆盖国民经济各行业的基本单位名录库、基础信息数据库和统计电子地理信息系统。此外，通过普查可以进一步夯实统计基础，完善国民经济核算制度，为加强和改善宏观调控、科学制定长期发展规划，提供科学准确的统计信息支持。确定调查目的，就是要明确在调查中要解决哪些问题，通过调查要取得什么样的资料，取得这些资料有什么用途等问题。衡量一个调查方案是否科学合理的标准主要看调查方案的设计是否能体现调查目的的要求，是否能够获取高质量的统计数据，以及是否具有较高的调查效率。

二、研究对象、调查对象和调查单位

明确调查目的之后，需要确定调查的范围以及向谁搜集统计数据资料的问题，即要明确研究对象、调查对象和调查单位三个问题。

研究对象是根据研究目的确定的，由某些性质上相同的许多个别事物组成的整体。例如，要研究某市居民的生活费收支情况，其研究对象就是该市的全部居民；要研究我国人口的情况，则研究对象是全国人口；要掌握我国农民的经济状况，则研究对象就是全国的农户。

调查对象是指根据调查目的确定的调查研究的总体或调查范围。调查单位是调查对象中的每一个单位，它是调查项目和调查内容的承担者或载体，也是我们收集数据、分析数据的基本单位。例如，人口普查中调查对象是全国人口；全国经济普查时调查对象是我国境内从事第二、第三产业活动的全部法人单位、产业活动单位和个体经营户。

三、调查项目和调查表

调查项目指的是调查单位所要登记的内容，一般表现为调查单位的各种信息，如在人口调查中，姓名、年龄、出生年月、民族、受教育程度、与户主关系等都是调查项目；在对企业经营状况调查时，企业的产值、产量、固定资产、企业性质等就是调查项目；在进行市场调查时，价格、销量、消费者满意度、消费者偏好等也是调查项目。确定了调查项目，就意味着明确了要向被调查者了解什么问题。一般，在确定调查项目时，需要注意以下几个问题。

(1) 调查项目的含义要明确。如果模糊不清晰，会使被调查者不能正确地理解调查者的想法，或使不同的调查者对问题有不同的理解，回答问题的标准不一致，导致调查所得到的数据无法使用。

(2) 设计调查项目时，既要考虑调查任务的需要，也要考虑调查的可行性问题。必要的内容不能遗漏，不必要的或不可能得到的项目没有必要列到调查项目中。

(3) 调查项目应尽可能做到有些项目之间相互关联，以便于资料相互对照，检查被调查者回答问题的逻辑性与准确性，且有利于分析研究现象发生变化的原因、条件和后果。

四、调查时间

调查时间是指调查资料、所登记数据资料的所属时间，即搜集什么时间上的数据。在调查方案中，必须对所搜集资料的所属时间做出具体规定。如果调查登记的是一段时期的资料，则应明确规定所登记的资料从何时起到何时止。如果调查登记的是某一时点上的资料，则应该明确规定统一的标准时点，例如，人口普查的调查方案中明确规定了调查的标准时点是调查年度的 11 月 1 日零点。

此外，设计调查方案时还需要确定调查工作的步骤及具体的调查工作的时间安排。

五、调查的组织实施

调查的组织实施主要包括调查的组织领导工作、调查机构的设置、调查人员的培训、调查经费的来源、调查方式方法的确定、调查工作的步骤、具体的时间安排等。对于一些重要的全面调查，往往需要组织试(预)调查，通过试调查发现问题并对调查方案进行调整。对于不同规模的调查，具体的组织实施会有很大差别，例如，在全国范围内组织实施一项全面调查，其组织实施是一个庞大复杂的工作，而进行企业内部、部门内部的调查，或进行小规模的市场调查，组织实施相对比较简单。

第三节　数据搜集的方法

一、数据的搜集方法

在确定调查方式或抽样方法的基础上，要根据调查对象的特点、调查的目的与要求，解决如何搜集到真实的统计数据资料的问题，即确定相应的数据收集方法。一般在实践中，常用的数据搜集的方法有访问法、观察法和报告法等。

(一)访问法

访问法是根据调查组织者事先所确定的调查事项，有计划地通过访谈询问的方式向被调查者提出问题，通过他们的回答来获得有关数据、资料的方法。在实际调查中，访问法有多种形式：

1. 面谈访问

这是调查者根据调查提纲、调查表或调查问卷直接访问被调查者，当面询问有关问题的一种调查方法，是一种使用频率非常高的方法。如采用街头拦访的方式，了解消费者对商品需求、购物习惯等方面的看法与行为，获取数据。这种方法由于有调查人员与被调查者的直接沟通，拒访率相对较低，调查表的回收率相对比较高，收集到的数据真实性比较强。但这种直接面对面询问的成本也相对比较高，调查结果、数据质量受到调查人员素质和工作态度的影响。

2. 电话访问

这种方法是由调查人员根据事先确定的抽样原则抽取样本，用电话联络的方式向被调查者询问，以收集有关数据资料的一种调查方法。电话调查的优点：时间短，速度快，节省经费，覆盖面广，可以对任何有电话的地区、单位和个人进行调查。但是，被调查者只限于有电

话和能通电话的地区的对象，电话提问受到时间的限制，问题数量不宜过多，与被调查者的沟通效果不如面谈，也不能出示调查说明、照片、图表等背景资料。

3. 电脑辅助调查

随着现代通信技术的发展，特别是电脑的应用，不仅数据分析可由计算机来完成，甚至整个调查过程，包括问卷设计和显示、样本设计、具体的调查、数据处理等都可以由电脑来控制或辅助完成。如电脑辅助电话调查是在使用电话调查时，借助于计算机来完成包括电话拨号、调查记录、数据处理等调查过程。

目前电脑辅助调查已经在我国的很多地区城市得到了较好的应用，普遍使用了电脑辅助电话调查系统，即 CATI 系统。该系统使电话调查更加便捷，也使调查数据的质量和调查的效率大大提高。

4. 座谈会法

座谈会法也称集体访谈法，它是将一组被调查者集中在调查现场，采用座谈会的方式针对调查的主题让被调查者发表意见，从而获取数据资料的方法。这种方式通常围绕研究主题，以一种比较自由的方式进行讨论。座谈会的方式决定了参加会议的人数不能太多，通常在十人以下，讨论的方式可以多种多样，目的是获取被调查者针对访谈问题的看法和建议，具体方式主要取决于会议组织者的习惯和爱好，对会议的主持者有较高的要求。

5. 个别深访

即深度访问，是一种一次只有一名受访者参加的特殊的定性研究，调查人员需要运用大量的追问技巧，尽可能让受访者自由发挥，表达他的想法和感受。其目的是不断深入受访者的思想当中，挖掘出表象、观点背后的动机。个别深访适用于研究较敏感、较隐秘不宜进行小组讨论的问题。

(二)报告法

报告法是由登记单位或报告单位根据原始记录和核算资料等，按照统一的表格和要求提供资料的方法。报告法具体的形式包括邮寄调查、日记调查、留置调查等。

1. 邮寄调查

这是调查人员将设计好的调查问卷或表格邮寄给被调查者，要求被调查者填妥后寄送、提交的一种调查方法。邮寄调查的优点：调查区域较广，调查成本较低，通过让被调查者用匿名的方式，可对某些敏感和隐私情况进行调查。其缺点是回收率相对比较低，信息反馈时间长，可能会影响资料的实效性。采用邮寄调查法的关键是选择好邮寄调查的对象。

2. 日记调查

它是指对那些需要连续调查的单位发放登记簿，由被调查者根据实际发生的事项逐日逐项记录，再由调查人员定期加以整理汇总的一种调查方法。例如我国城市职工家计调查就是采用这种方法，由被调查户登记家庭生活费收支等情况。日记调查的优点是：能使调查双方建立长期联系，回收率较高。同时由于每天记录，能比较详细、确切地反映被调查者的经济活动情况，便于统计资料在不同时间和地区做各种对比分析。其缺点是登记、记账工作量很大，许多主、客观因素都会影响记账的连续性和准确性。为了克服这些缺点，在实际工作中应采取加强宣传和奖励并重的做法，加强辅导和资料审核、汇总的工作，必要时需要采用国际上比较流行的样本轮换办法。

3. 留置调查

它是指调查人员将调查问卷送交被调查者，说明调查意图和填写要求，由被调查者自行填写回答，再由调查者按约定日期收回的一种调查方法。留置调查是介于面谈和邮寄调查之间的一种方法，此法可弥补当面提问时容易受时间限制、被调查者考虑问题不成熟等缺点，又可克服邮寄调查回收率低的不足。它的缺点是调查地域、范围受到一定的限制，调查费用相对较高，调查时间也较长。

(三)观察法

它是指调查者通过直接观察、跟踪等方式，记录被调查者的情况来收集资料的一种调查方法。观察法不同于日常生活中的“观察”，它具有目的性、计划性和系统性，而且要求观察者对所观察到的事实做出实质性的结论。例如，在试销某种新产品时，需要观察消费者对该产品的反应，通过受过专门训练的观察人员或隐蔽的录像机，记录下有多少人走过售货架，多少人停下来，细心地观看、选择、购买或者又重新放回，他们的性别、年龄怎样，都有些什么表情和动作等情况。例如在城市集贸市场调查中，对市场上农副产品的上市量、成交量和成交价格等情况进行观察。

观察法的最大优点是它的直观性和可靠性。它简便、易行、灵活性强，可随时随地进行调查。但若观察不够深入、具体，只能说明事实的发生，而不能进一步地说明发生的原因和动机，也有些调查无法采用观察法，如对历史资料的收集和对居民手持现金数量的调查等等。此外，该方法调查时间较长，耗费人力、物力较多，受时间、空间和经费限制较大。

上述各种收集数据的方法都有各自的优缺点，而这些优缺点都是相对而言的。在实际工作中，由于调查目的不同，调查的侧重点不同，调查对象的具体情况也不同，就需要对这些方法进行比较分析，择优选用适合的调查方法。

选择具体的方法时，需要综合考虑各种方法的调查范围、调查对象可控性、影响回答的因素、回收率、回答速度、回答质量和费用等，以便能够选择出最适合的方法。

二、问卷的设计

(一)问卷的概念和作用

问卷是调查中使用最为普遍的，用于收集数据的一种表述调查项目的形式。掌握问卷设计技术，对于取得能够满足调查目的的统计数据资料具有十分重要的意义。

问卷是指对询问的问题或提纲按要求填选答案的调查表。问卷不但有利于调查内容、调查项目的系统化、标准化，便于对所取得的数据进行统计处理和定量分析，而且还可以节省调查时间和提高工作效率。由于问卷中的许多问题都已经给出可供选择的答案，易被调查者所接受，因而问卷调查已成为收集数据的重要手段。问卷设计得是否科学、可行，在很大程度上决定了数据收集工作的成败和获取数据的质量。因此设计一份科学、完善的问卷，是调查工作成功的重要保证。

采用问卷调查可以使调查研究内容规范化，提出的问题与答案标准化，也可以使调查研究程序化。调查访问按问卷规定的提问和回答次序进行，有利于调查的科学化，便于提高数据的可靠性和分析数据的正确性。

(二)问卷的基本内容与格式

一份完整的调查问卷通常包括以下内容。

1. 问卷的标题

问卷的标题要求能概括地说明调查研究的主题，使被调查者对所要回答的问题有一个大致的了解。标题的设计应简明扼要，易于引起被调查者的兴趣。

2. 问卷说明(前言)

问卷说明是给被调查者的简短信函，一般写在问卷的首项或封面上，其主要内容包括：(1)自我介绍，即调查人员的身份说明；(2)调查的目的及重要性，以引起被调查者的重视与关注；(3)相关问题的承诺，如对被调查者的相关信息和观点承诺保密，以解除被调查者的顾虑；(4)请求合作并表示谢意等。

3. 被调查者基本情况，即背景信息

这是指被调查者的一些主要特征，如性别、年龄、民族、文化程度、职业、所在地区等等。其目的主要有：(1)调查工作完成后，可以利用这些信息考察样本单位的分布是否合理；(2)便于数据处理时的交叉分组，即反映不同人群的观点、状况，以从调查中发现规律与存在的问题等。

4. 调查的主体内容

这是调查者所要了解的基本内容，也是调查问卷中最重要的部分。它主要是以提问的形式提供给被调查者，这部分内容的设计好坏会直接影响到整个调查的效果。

5. 编码

它是将问卷中的调查项目变成数字、用自然数给各种答案编上号码的工作过程，其目的是便于分类整理，易于进行计算机处理和统计分析。其主要作用如下。(1)对没有明确数量化的研究对象起数量化的作用，如要了解被调查者对目前的收入水平的满意程度，满意程度本身是一个模糊的概念，调查中我们需要将其程度测量出来，才能进行汇总、分析，这样就可以将可能的结果分为五种情况：很满意、满意、一般、不太满意、不满意，并用数字 1、2、3、4、5 作为其相应的编码。(2)对已经数量化的研究对象起归纳分类的作用等。

6. 作业证明的记载

在调查问卷的最后，附上调查员的姓名、访问日期、时间等。如果有必要和可能，还可加上被调查者的姓名、电话等，以便于审核和进一步追踪调查。

(三)问卷设计中的询问技术

问题是问卷的核心，设计问卷时，必须仔细研究问题的类别和提问的方法。

1. 问题的主要类型及询问方式

第一，根据所提问题的类型不同，可分为直接性的问题、间接性的问题和假设性的问题。一般对于像性别、对某问题的一般性看法与建议等对被调查者来讲不敏感、一般不会有顾忌的问题可采用直接提问的方式。而对于那些可能涉及被调查者个人秘密或隐私等不愿意直接回答的问题，询问时可采用间接提问的方式，比如有关收入的问题，如果直接问，很多人不愿意回答或即使回答也不一定是真实的，这时就可以考虑将收入分为若干组，由被调查者选择自己的收入所属组别即可，或也可以问其家庭食品支出占消费支出的大概比重，由此去推断其收入水平。对于涉及被调查者对某些问题的看法、立场或对未来的想法、期盼、建议等类似的问题，询问时可采用假设性的提问方式。例如，了解消费者消费行为、偏好方面

的问题，可以问：“如果在购买汽车和住房中您只能而且必须选择一种，您可能选择何种？”

第二，根据对问题的作答方式不同，问卷中的问题可分为开放性问题、封闭性问题等。所谓开放性问题，是指在问卷中不列出所有可能的答案选项，而由被调查者自由作答。例如，“您对进一步提高公共满意度有何建议？”开放性问题的主要优点在于被调查者可以充分地按自己的想法与方式回答问题和发表意见，不受问卷设计者思维或想法的干扰，也不受问卷中有限的选项的限制，有利于充分发挥被调查者的主动性和想象力。因此开放性问题所得的信息与资料往往比较丰富、具体且信息量大，特别适合于询问那些潜在的答案有很多，或者答案比较复杂，或者组织者尚未弄清各种可能答案的问题，尤其是想要探讨建设性的意见和建议时，可采用开放性的问题。当然，对于开放性问题，由于回答者提供答案的想法和角度不同，语言表述方式和特点不同，因此在对答案进行统计处理和分析时，工作量相对较大，同时还可能因为分析者不能完全理解回答者的意思而导致调查偏差。此外，由于时间关系或缺乏思考，事先没有充分准备，被调查者往往会放弃回答或答非所问，使问卷的回收率(或单题回收率)和有效率降低。因此问卷设计中，开放性问题不宜过多。

封闭性问题是指在所提问题即题干后面，列出事先设计好的所有可能的几种答案，被调查者根据自己的具体情况从中选定最合适的或者认同的答案选项作为回答。

例如，您家庭的目前收支情况总体上是：

A．较多的节余　B．略有节余　C．收支基本平衡　D．入不敷出

封闭性的问题由于答案标准化，因此回答方便，易于进行各种统计处理和分析，有利于提高问卷的回收率和有效率。当然，设置封闭性的问题，要求所有选项是完备的，且各个选项之间应该是互斥的，否则一旦设计有缺陷，如被调查者还有其他的答案中没有涉及的想法与情况，而回答者只能在规定的范围内选择、回答，就可能无法正确回答问题，从而影响调查的质量。

为了避免封闭性问题的上述可能存在的缺陷，在设计答案选项时，可以将封闭性与开放性问题结合，形成半开放、半封闭性的问题，即主体选项是封闭性的，同时答案中另设计一个选项为“其他”，当被调查者选择“其他”选项时，要求其进一步注明其内容。如：

您目前最迫切需要解决的问题是(　　)

A．购买住房　B．子女入学及教育　C．提高收入　D．医疗保障

E．就业　F．带薪休假　G．其他(请注明)_____

第三，根据所提问题的内容，可分为事实性问题、行为性问题、动机性问题和态度性问题。

所谓事实性问题是指要求被调查者回答一些相关事实，其主要目的是获得反映客观事实的资料，因此问题的涵义必须清楚，使被调查者容易理解并易于回答。通常，在一份问卷中涉及的被调查者的个人信息资料，如职业、性别、年龄、家庭状况、教育程度等，这些问题均为事实性问题。

行为性问题是对被调查者的行为活动进行调查。例如：

以下社区文化设施中，您最经常去的是(限选三项)：

A．图书馆　B．文化站　C．社区文化室　D．文化广场

E．影剧院　F．博物馆　G．健身运动场所　H．其他(请注明)_____

设置动机性问题是为了了解被调查者行为的原因或动机。应特别注意的是，在提出动机

性问题时，应注意人们的行为可以是由有意识的动机产生的，也可以是由半意识动机或无意识动机产生的。对于前者，有时被调查者会因为种种原因不愿意真实回答；对于后两者，因被调查者对自己的动机不十分清楚，从而也会造成回答的困难和回答的有效性偏低。

设置态度性的问题是为了了解被调查者对某一事物的态度、评价、意见等而提出的问题。例如，调查对某商品的满意度时提出“你对该商品的功能是否满意？”的问题就是态度性的问题。

问卷中提问的方式可分为各种不同的类型，在实际应用中应注意结合各类问题的特点恰当地选用，当然实际调查时，上面几种类型的问题也经常是结合在一起使用的，如前面提到的半开放、半封闭性的结构性问题。同样，事实性问题除了采取直接提问的方式以外，对于回答者不愿意回答的问题，也可以选择间接性的提问方式。

2. 设计问题问句时应注意的几个问题

对问题问句的设计，尤其是封闭性问题在设计问句时，总体上要求问句表达一定要简明、生动，注意概念的准确性，避免提似是而非、模模糊糊的问题。

第一，避免提笼统的、抽象的问题。当调查对象为一般的、普通的被调查者时，应尽量避免提出过于专业化的问题，这样的问题容易造成被调查者对问题理解上的困难，不易回答，且有时对实际调查工作无指导意义。同时，提出的问题要具体，不要过于笼统，例如，“您对***超市的印象如何”，这样的问题就过于笼统，被调查者不容易把握提问者问的是对哪方面的印象，因而很难达到预期效果，此问题的提出可具体一些，如“您认为***超市商品品种是否齐全”“您认为***超市营业时间是否合适”“您认为***超市服务态度怎么样”等。

第二，避免用不确切的词语。例如“普通”、“经常”以及一些形容词等，因为对这些词语，不同的人可能会有不同的理解，如：“您是否经常到***超市购物？”在回答这个问题时，被调查者不好把握“经常”的概念指的是一周至少一次、一个月至少一次还是每天一次，提问时，可改写为“您上周到***超市几次”，或在“经常”后面注明“指每周至少一次”等语句。

第三，避免使用含糊不清的句子。例如，“您这次是出门旅游，还是休息？”出门旅游本身就是休息的一种形式，它和休息并不是互斥的关系。如果改写为“您在下个休息日是出门旅游，还是在家休息”则会好些，不会让被调查者在词语方面感到不知所措。

第四，避免引导性的提问。如果提出的问题不是折中的，而是暗含着调查者的观点和看法，则易使回答者跟着这种倾向回答，这种提问就是引导性的问题。如“大家都认为***是一个优秀的企业家，您的看法如何？”引导性的问题会导致两种不良的后果：一是被调查者不加思考就同意题干中暗示的结论；二是由于引导性问题的提出使被调查者心理上产生某种顺向效应或逆反心理，从而导致最终的选择与心理感受不一致的结果。因此，这种提问在调查中一定要避免。

第五，避免提断定性的问题。例如：“您一天抽多少支烟？”“您认为产品质量不好的最主要原因是什么？”这种问题为断定性的问题，如果被调查者根本就不吸烟，或认为产品质量很好，就会造成无法回答的现象。正确的处理方法是就此问题可加上一个过滤性的问题，即“您抽烟吗？”然后再问：“如果抽烟，您一天抽多少支烟？”当然也可以从过滤性的问题采用跳问的方式，如果选择抽烟，直接跳问到“您一天抽多少支烟？”这个问题上。

第六，避免提出令被调查者难堪或被调查者禁忌和敏感的问题，如涉及各地风俗及民俗习惯中忌讳的问题，以及关系个人利害关系和个人隐私的问题等。

第七，避免提出复合问题。所谓复合问题是指在一个问题中包含多个调查项目或内容，如：“你对***电视机的清晰度和色彩满意吗？”对于这样一个封闭性的问题，被调查者很多时候无法选择，如果被调查者只对其中一个方面满意，那么他回答“是”或回答“否”都会有问题，为避免此结果的出现，设计问句时要分离语句中的提问部分，使得一个问题只问一个调查项目。

(四)封闭性问题中答案的设计技术

答案设计，也同样是问卷设计的重要组成部分，封闭性问题中的答案设计显得尤为重要，需要反复的、多方面的、周密而细致的推敲。总体上，答案选项设计的基本要求是穷尽、完备且互斥，所谓穷尽、完备是指答案设计中的选项应包括所有可能出现的情况，不至于被调查者因为找不到合适的选项而放弃回答，而互斥则要求各选项之间需要互斥，即互不包容、互不重叠，否则可能会导致被调查者做出重复内容的双重选择，影响调查效果。答案选项设计的基本方式、方法与要求如下。

1. 二选一

这是指所提出的问题只有两种答案，这两种答案是对立的，被调查者只需要从两个选项中选择其中的一个即可，不可能有更多的选择。如：“您的性别是”，答案只能是“男”或“女”；“您是否看过***节目”，答案只能是“是”或“否”。这种方法的优点是：易于理解，可迅速得到明确的答案，但是该类问题的信息量有限，所掌握的信息只是最表层的，因而只适用于较为简单的、事实性的问题。

2. 多项选择

这是指所提出的问题事先预备好两个以上(不含两个)的答案选项，被调查者可在各选项中选择其中一项或几项回答。

例如：您认为目前对北京城市环境威胁最大的因素有(　　)(限选三项)

(1)塑料包装等白色污染物　(2)废电池等电子垃圾　(3)噪音污染
(4)机动车污染物排放　(5)工地扬尘　(6)水污染
(7)生活垃圾　(8)其他(请注明)_____

应该注意的是：第一，由于所设答案选项不一定能表达出被调查者所有的看法，所以在问题的最后通常可设“其他”选项，以便使被调查者表达自己的看法；第二，多项选择中，由于选项较多，选项的顺序可能会影响被调查者的选择，从而使调查结果发生偏差。为避免这一问题的出现，设计选项顺序时，可以是随机排列，也可以根据选项字数从少到多进行排列。

3. 排序法

这是指列出若干答案选项，由被调查者按选项的重要程度决定其先后顺序的方法。排序时，可以将全部选项排序，也可只对其中最重要的几项(可事先确定项数)进行排序。

例如：您认为目前对北京城市环境威胁最大的主要影响因素有(请按程度大小顺序排列)_____

(1)塑料包装等白色污染物　(2)废电池等电子垃圾　(3)噪音污染
(4)机动车污染物排放　(5)工地扬尘　(6)水污染
(7)生活垃圾　(8)其他(请注明)_____

4. 两两比较法

这是把若干可比较的事物整理成两两对比的形式，要求被调查者进行比较并做出肯定回答的方法。例如：

请比较下列每一对文化产业园区，哪一个是您更喜欢的？(每一对中只选一个划√)

□ 中国(怀柔)影视基地　　□ 798 艺术区
□ 798 艺术区　　□ 潘家园古玩艺术品交易园区
□ 潘家园古玩艺术品交易园区　　□ 宋庄原创艺术与卡通产业区
□ 宋庄原创艺术与卡通产业区　　□ 琉璃厂历史文化产业园区
□ 琉璃厂历史文化产业园区　　□ 北京(房山)历史文化旅游区
□ 北京(房山)历史文化旅游区　　□ 前门传统产业区
……

比较法适用于对质量和效用等问题做出评价，应用比较法要考虑被调查者对问题中的项目是否比较熟悉，否则将会导致空项的发生。

5. 避免问题与答案不一致

所提问题与所设答案应做到一致，避免所答非所问。例如，问题是“您经常看哪个栏目的电视节目”。

若设计的答案选项是：

A．经济生活　　B．道德观察　　C．新闻联播　　D．其他(请注明)_____
E．经常看　　F．偶尔看　　G．根本不看

就出现了问题与答案选项不一致的现象，会出现多余或矛盾的选择。

(五)问卷的编排设计和要求

设计好问卷的询问方式和相应的答案后，下一项工作就是对问卷进行编排。对问卷所设计的问题进行编排，一般要注意以下几个问题。

1. 问卷的顺序

在设计问卷时，要注意问题的排列顺序。问题排序的基本要求是使问卷条理清楚、顺理成章，便于被调查者回答与合作，从而提高回答问题的效果。一般问题的顺序：

第一，容易回答的问题放在前面，较难回答的问题放在中间，敏感性的问题放在后面。

第二，封闭性的问题放在前面，开放性的问题放在后面。这是由于封闭性的问题有现成的答案，较易回答，而开放性的问题需要被调查者花费一些时间考虑并填写，放在前面容易让人产生畏难情绪，从而不利于调查的顺利进行。

第三，要注意问题之间的逻辑顺序，如可按时间顺序、按类别顺序等进行排列。

2. 问题的衔接

问卷中的各种问题应很好地衔接起来，使符合某种回答条件和不符合某种回答条件的被调查者均可快捷方便地找到其应该回答的问题和选项。

例如：

您家有汽车吗？

A 有□　　如果有，您家的汽车是(1)国产　(2)进口　(3)合资

B.无□

有时，如果连续几个问题都适合于具有某种条件的被调查者，设计时可以采用跳答的方法来解决。例如：

Q11．您看过《**导报》这份报纸吗？

A．经常看　　B．偶尔看　　C．从没看过(跳转到问题 22)

Q12．您对《**导报》的总体评价是：

A．满意　　B．一般　　C．不满意

Q13．……

第四节　统计数据的质量与类型

一、统计数据的质量

统计活动的全过程都是在与数据打交道，从数据的收集开始到后续的整理、分析，每个环节都少不了数据。统计数据的质量，直接影响到分析的结果与结论，因而对统计数据质量的控制，应该贯穿统计活动的所有过程，但其中以调查阶段的数据质量最为重要。

当我们面对一组统计数据时，往往首先考虑的问题是这些数据是否准确、是否真实可靠。数据的准确性、真实性与可靠性是统计的生命。如果研究、分析、描述问题所依据的数据存在问题，那我们是不可能得到正确的结论的。

评价数据质量的标准或对数据的基本要求是准确、及时、全面、系统。而其中数据的准确性，我们往往可以从统计数据误差的角度进行评价。为使数据质量检查有一个定量描述的依据，我们把统计数据质量的准确性解释为：获得的观察值与客观现象实际数量水平之间的离差。离差大说明数据的准确性差，即数据的质量较差，离差小说明数据的准确性高，即数据的质量好。

二、统计数据的误差

进行数据观测与计量的过程中，很多因素都会使数据产生错误或误差。总体看，统计数据的误差包括抽样误差和非抽样误差两类。

(一)抽样误差

利用概率抽样调查收集样本数据的主要目的是对总体特征进行估计与推断，但是由于样本是随机抽取的，因此抽中不同的样本，就会有不同的估计值。即使是所收集的样本数据完全正确可靠，它和总体真值之间也往往会存在差异，这种差异是样本对总体的代表性导致的，所以称为抽样误差。如，在对市场上某品牌产品进行质量抽检时，假设抽检的产品采用的是随机抽样的方式从市场上所有该品牌产品中抽取的，在检测无误的情况，如果样本产品合格率为 96%，也并不代表该品牌产品的总体合格率就是 96%，一般我们认为可能是 96%左右。因为再抽取一次，抽检的产品和上一次可能不完全一样，其合格率未必正好等于 96%，可能是 97%，也可能是 95%。正常情况下样本合格率和总体真正的合格率多多少少会有点差异，即抽样误差。抽样误差并非错误的误差，而是由于样本的代表性产生的误差。

如果我们要估计总体的均值为 μ，随机抽样得到样本的均值为 $\bar{x}$，则 $\bar{x}-\mu$ 就定义为抽样误差。但是在实际的抽样中，由于总体均值 μ 是未知的，因而实际的抽样误差 $\bar{x}-\mu$ 也是无法知道的。但是根据抽样原理我们可以计算出所有可能样本均值与总体均值之间的平均离差，我们称其为抽样平均误差。由于抽样误差有正有负，相互抵消后其平均数为零，所以采用计算标准差的方法来计算平均误差，因而称为抽样标准误差，后面简称为抽样误差，其定义公式为 $\sqrt{E(\bar{x}-\mu)^2}$。它可以根据不同的抽样方法的公式计算，抽样标准误差越小，说明该抽样方案的精度越高，它可以作为控制抽样误差的重要依据。

一般在公布抽样调查的结果时，要说明抽样误差的大小，不管是比例、均值还是其他形式。抽样误差告诉我们的是，样本离总体的实际值可能有多远，其具体的计算和应用将在后面章节中介绍。

(二)非抽样误差

在实际的数据观测与计量中，除了由于样本的随机性导致的代表性抽样误差外，还会有由于其他因素产生的误差，如有的来自于调查员的疏忽导致的错误；有的来自被调查者有意或无意的虚报或瞒报；也有来自指标设计、问卷设计时存在问题而导致被调查者对问题的错误理解；等等。这些误差统称为非抽样误差。几种比较典型的导致非抽样误差的因素主要有。

(1)未响应导致的误差，也称为无回答误差。未响应误差是指在调查中被抽中的调查者拒绝回答所提出的问题而导致的误差。无回答一般有两种情况，一是有意无回答，二是无意无回答。对于后者我们可以用替代样本单元的数据替换，不会导致系统偏差，而有意无回答则可能会导致总体数据估计的偏差。有意无回答现象对总体数据估计推断可能会产生以下影响。

第一，由于无回答现象的出现而使有效的样本量减少，从而使抽样误差增大，达不到原抽样设计时对调查精度的要求。样本量的大小会直接影响到抽样误差，这将在后面的章节中介绍。

第二，由于无回答问题而带来的估计量的误差，而且这种误差并不会由于样本量的增大而减少。无回答误差的产生主要来自于回答的人与无回答人之间的态度或标志值的差异。误差的大小取决于回答者和无回答者之间对问题看法的差距，以及无回答的问卷在所有问卷中所占的比例。

例如，某单位对其职工的年个人收入进行调查，假设抽取了 100 人进行了调查，调查时这 100 人均在单位，但只有 80 人回答了该问题，计算结果是其平均年收入为 44 870 元，如果没有作答的 20 人与这 80 人的水平相近，则 44 870 元就可作为推断全部职工个人年平均收入的依据。但是，如果随后的进一步调查发现，没有作答的 20 人，大多数在外有兼职的收入，计算起来这 20 人的个人年平均收入为 62 800 元，这样，被抽中的 100 人的平均收入应为(44 780×80+62 800×20)/100=48 384 元，高于 80 人的平均水平 44 870 元。

再如，假定计划调查 1200 人，却只有 1000 人接受了调查，这意味着缺少了 200 人的样本数据。在 1000 个回答者中，我们发现有 600 人即 60%的人赞成某事物而其余人反对。如果我们假定另外 200 人也赞成，那么在 1200 人中就有 800 人赞成，比例为 67%。但另一方面，如果我们假定那 200 人反对，那么 1200 人中只有 600 人赞成，比例为 50%。可见，仅仅由于部分人未回答而产生的未响应误差，观测样本中 60%的赞成比例有可能实际只是 50%～67%之间的一个随机数，这就可能对我们的研究结论造成很大的影响。

在实际调查时，高拒绝率是调查者、研究者可能经常遇到的一个很大的问题。出现这样的现象，我们要对拒绝回答的原因进行分析，还要分析回答者和未回答者之间的区别。相对于回答者来说，未回答者有什么不同和特点呢？如果他们回答，他们的回答会怎样影响结果呢？

一些经验表明，在大部分情况下，未回答者和回答者并无多大的差别。如果我们开始时有一个高的回答率，那么可假定未回答者也以同样的比例做出回答。但是如果未回答率比较高，那么未响应误差有可能对结果产生明显的影响。

(2)抽样框导致的误差。所谓抽样框就是在抽样时用以从中抽取样本单位、包括研究对

象所有单位的一个名单或框架。抽样框的形式可以是研究对象中所有单位的名单形式，也可以是其他的形式，如地图，甚至是一个概念。如职工的名单、地区或区域地图等，若想了解观众对新上映的大片的评价，我们找不到所有观看影片的观众名单，也用不了地图，这时抽样框则可定义为“该电影上映期间观看该影片的观众”，这就是一个概念框。

根据研究目的所确定的理想总体称为目标总体，抽样框所代表的总体通常称为抽样总体。从理论上讲，这两个总体应该是一致的。然而在实际调查中，我们会发现，与目标总体完全一致的抽样总体不一定是找得到的，或要取得这样的抽样框需要过多的无法承受的人力、物力和财力，这时，我们通常只能用一个接近目标总体并且容易取得和便于操作的抽样总体来代替。由于样本是从抽样总体中抽取出来的，因而用样本数据进行推断的应该是抽样总体的指标，和目标总体的指标多少会有一定的差距，即出现了误差。我们把由于抽样框与目标总体不一致而导致的误差称为抽样框误差。

由于抽样框导致的偏差甚至是调查的失败，最典型的例子是 1936 年美国总统竞选时《文学摘要》杂志所做的民意测验。《文学摘要》是美国一家著名的杂志社，为了预测 1936 年美国总统竞选的结果，竞选前发出近 1000 万张问卷，调查并预测共和党候选人兰登和民主党候选人罗斯福谁将获胜。这项调查的样本量应该说是足够大，但调查推断结果和实际的选举结果完全相反，下表列出了调查数据与最终选举的结果：

候选人	得票率的调查结果(%)	实际选举的得票率(%)
罗斯福	43	62
兰登	57	38

事后分析导致调查推断失败的原因之一就是抽样框存在问题。因为该杂志当时是以电话簿作为抽样框进行抽样的，而当时在美国电话还不普及，只有比较有钱的人才装得起电话，而有钱人中较多的是共和党人，因而在民意测验中主要反映的是这部分人的意见，而实际选举中共和党人只是其中的一部分，而不是全部，这就势必造成调查的误差。

抽样框与目标总体的不一致有多种情况，最常见的是抽样框中丢失目标总体单位和包含非目标总体单位。丢失目标总体单位，即抽样框有遗漏，如果丢失的总体单位的数据与未丢失的总体单位数据之间有差异，就会在估计总体平均数据时产生偏差，这种偏差可能偏大，也可能偏小，但在估计总体总值时，往往由于丢失总体单位而使估计值偏低。至于包含非目标总体单位的情况，若抽中这些单位时可以进行鉴别而剔除，则对估计总体均值的影响比较小，但在估计总体总值时会由于总体单位数中包含了非目标总体单位而偏大。当然，也有的抽样框既丢失部分目标总体单位，同时也包含了一些非目标总体单位。实际调查中，这种误差也是经常存在的，需要引起我们的高度关注。

抽样误差与非抽样误差的性质不同，前者是一种代表性误差，在随机抽样中有时抽中偏小的单位，有时则抽中偏大的单位，因而这种误差没有系统性的方向，随着样本量的增大，其抽样误差由于相互抵消而缩小。而非抽样误差则不同，它往往有系统性，根据不同的情况会偏向某一方向。例如，有些反映成绩的数据可能会普遍偏高，而另一些指标，尤其是逆指标(即越小越高的指标)则普遍偏低，此外非抽样误差不会因样本量的增大而减小。

三、统计数据质量的检查与要求

对统计数据质量的检查，主要有两大类方法：后验方法和抽样检查方法。

（一）后验方法

后验方法是在调查工作完成后，不用亲临调查现场，而是通过对数据的逻辑关系分析、计算比较，以及将调查数据与独立来源的资料对比，对调查数据的质量进行分析。所谓逻辑关系分析，就是把调查数据与人们普遍接受的对现象某些特征或关系的看法进行比较，以判断有无矛盾的地方。如对人口普查数据，我们可以审查年龄和婚姻状况之间的逻辑关系，若年龄15 岁的人，其婚姻状况为已婚，则该数据很可能是有问题的，这是一种逻辑上的判断。当然，对有些数据可能需要进行相关的计算比较，如利用数据之间的平衡关系，看其是否平衡，以此分析数据的质量。此外，由于现象之间客观上存在一定的量值范围和比例关系，根据这种量值的范围和比例关系，按照规定检查的参照标准，检查数据是否有问题。如一般收入水平越高，其食品支出所占比重则应越低，据此通过一定的标准来大致分析数据的质量。

使用后验方法分析数据质量时，需要注意其适用条件，后验方法常常适用于研究具有某种内在规律性特征的现象，对那些调查特征变化无常的现象或我们不知道其规律的现象，后验方法的效果不会很好。同时后验方法仅适用于对最后调查结果的检查，不能用于单项数据误差的评估，因而对于改进调查过程中的收集方法有一定的作用，但作用有限。

（二）抽样检查方法

抽样检查是指在一次调查之后，在尽量短的时间内再从这些被调查单位中抽取一定数量的单位组成样本重新进行调查，将两次调查的结果进行比较，以分析调查数据的质量，并对所收集的数据进行修正。如在人口普查后，就是通过抽样调查的方式对普查数据进行质量评估和数据的修正。

通过抽样方法进行数据质量检查所得的结论，是完全根据样本资料得出的，因此不管有没有相关的统计数据可依，无论过去是否做过类似的调查，都不会影响已收集到的数据质量的评估。抽样检查数据质量应用灵活，适用于各种场合，也适用于调查数据中各部分的检查。但是也应该注意，进行抽样检查必须在一次调查之后不久就要进行，以免调查的对象及调查环境发生较大的变化而导致的评估无效。同时由于样本数据是检查的标准，因而样本抽取、数据的收集与核查，都需要安排专业的统计人员进行。此外，样本单位的确定要考虑到随机性的要求，而且在考虑预算费用约束的情况下，要保证有合理的样本量。

四、数据的类型

统计数据是对研究对象及其活动的过程或结果进行观测、计量的结果，例如，对经济活动总量的计量可以得到国内生产总值的数据，对人口性别的计量可以得到男女性别比的数据，对城镇居民收入与消费的调查可以得到城镇居民的收入水平和消费支出的数据，等等。

（一）数据的计量尺度

对不同的事物进行计量或测度的精确程度是不同的。按照对事物计量的精确程度，可将所采用的计量尺度由粗略到精确分为几个不同的层次。

1. 定类尺度

定类尺度是最粗略、计量层次最低的计量尺度。它的主要特征是用文字、数字代码和其他符号对事物的特征进行表示，利用定类尺度测量的数据仅仅能够进行分类或分组。我们通常把这些仅建立在对事物进行分类分组的基础上的计量尺度称为定类尺度。如对人口性别、

民族、籍贯、婚姻状况的统计测量，对企业经济性质的划分等作归类统计的测量均属于定类尺度。定类尺度只测量事物之间的类别差，各类别之间无法区分优劣或大小。

由于定类尺度只能区分事物是同类或不同类的，因而使用定类尺度计量的数据不能直接进行数学运算。当处理数据时，为了便于计算机识别和信息的传输，对于定类尺度计量的数据，我们可以给每一个类别赋予一个数字代码，如在人口性别中用“1”表示男，用“0”表示女，然而要注意的是，这些数字只是给不同类别的一个代码，并不意味着这些数字可以进行任何的数学运算。

在使用定类尺度对事物进行分类时，必须符合穷尽和互斥的要求，即在进行分类时，必须保证所有个体都能归属某一类别，不能有遗漏，而且只能归属于一个类别，不能在其他类别中重复出现。

2. 定序尺度

定序尺度又称顺序尺度，它是对事物的等级或顺序的一种测度。该尺度不仅可以将事物分成不同的类别，而且还可以确定这些类别的优劣和等级顺序。如对城镇居民进行调查，了解居民对公众服务的满意度，满意程度可分为很满意、比较满意、一般、不太满意和不满意；被调查者的学历可分为小学及以下、中学、高中、大专、大本和研究生；产品按其品质分成一等、二等、三等，等等，都属于定序尺度。定序尺度的计量结果虽然也表现为类别，但这些类别之间是可以比较并排序的，其对事物的计量要比定类尺度精确一些，但它也只是测量了类别之间的顺序，不能测量出类别之间的准确差值。该尺度可以分类，也可以比较优劣、好坏，但不能进行加、减、乘、除等数学运算。

3. 定距尺度

定距尺度也称间隔尺度，它不仅具有定序尺度的所有特征，即能将事物区分为不同类型并进行排序，而且可以准确地指出类别之间的差距是多少。定距尺度是对事物类别或次序之间距离的度量，通常使用自然或度量衡单位作为计量尺度，其计量结果表现为数值。例如，考试成绩用百分制度量，温度用摄氏度或华氏来度量，等等。由于这种尺度的每一单位间隔都是相等的，只要给出一个度量单位，就可以准确地指出两个计数之间的差值，因而其计量结果不仅可以进行分类和比较，也可以进行加、减的数学运算。

4. 定比尺度

定比尺度是最高级别的测量尺度，其计量结果和定距尺度一样也表示为数值。但是与定距尺度不同的是，定比尺度拥有一个绝对零点的测量原点，而定距尺度却没有这样的原点。因而定距尺度只能进行加、减运算，而定比尺度不仅可以进行加、减运算，也可以进行乘、除运算。如工资是定比尺度，其绝对零点是零元，即没有。若一个人的月工资是 6000 元，另一个人的月工资是 3000 元，可以得出前者工资比后者多 3000 元，是后者的两倍。再如，温度是定距尺度，其不存在绝对零点，即使温度为零，也不能说没有温度，我们可以说 30 度比 15 度高出 15 度，但不能说 30 度比 15 度热一倍。

采用不同的计量尺度可以得到不同类型的统计数据，而不同类型的统计数据又适用于不同的统计分析方法。定类、定序、定距、定比四种计量尺度对事物的测量层次是由低级到高级逐步递进的，高层次的计量尺度具有低层次的计量尺度的特征，我们可以将高层次的计量尺度的测量结果转化为低层次的计量尺度的测量结果，如将温度的定距尺度测量结果转化为很冷、较冷、舒适、较热、很热的定序列尺度的测量结果。四种计量尺度的数学特征是：(1)定比尺度可以分类、排序、测量距离和计算比值；(2)定距尺度可以分类、排序、

测量距离，不能计算比值；(3)定序尺度可以分类、排序，不能测量距离和计算比值；(4)定类尺度只能进行分类。

(二)数据的类型

所谓数据是对现象进行计量的结果，如对气温的测量、对经济活动总量的计量等。我们可以从不同的角度对统计数据进行分类。

1. 分类数据、顺序数据和数值型数据

按照所采用的计量尺度，统计数据可以分为分类数据、顺序数据和数值型数据。

分类数据是对事物进行分类的结果，描述事物的品质特征。例如，人的性别：“男、女”；职业：“教师、公务员”等；籍贯：“北京、天津、河北”等；企业经济性质：“国有、股份制”等都是分类型数据。分类数据的表现通常是用文字来表述的。当然，为了便于统计处理，对于分类数据我们也可以给不同类别赋予不同的数值，如用“1”表示男，用“0”表示女，但这并不影响其性质和数学特征。

顺序数据也是对事物进行分类的结果，但这些类别之间可以进行优劣、好坏的排序。如很满意、较满意、一般、不太满意和不满意，就属于顺序数据。与分类数据相同，顺序数据的表现也通常是用文字来表述的，但为了便于统计处理，也可以用 1、2、3、4、5 数字代码来代表上面的文字表述。

数值型数据是用定比尺度和定距尺度计量、测量的结果，用以说明事物的数量特征。如人的年龄 20 岁、企业的人数 500 人、企业的产值 1300 万元、地区生产总值 2367 亿元等。数值型数据表现为具体的数值，这也是统计学研究中最多的数据类型。

2. 截面数据、时间数列数据和平行数据

数据按照被描述的对象与时间的关系分为截面数据、时间数列数据和平行数据。

截面数据描述的是事物在某一时刻、时点或时期上的特征。例如某年全国各省市、自治区的地区生产总值数据(见表 2-1)。

表 2-1　某年各省市地区生产总值　　单位：亿元

地　区	地区生产总值	地　区	地区生产总值	地　区	地区生产总值
北　京	9353.32	安　徽	7364.18	四　川	10 505.30
天　津	5050.40	福　建	9249.13	贵　州	2741.90
河　北	13 709.50	江　西	5500.25	云　南	4741.31
山　西	5733.35	山　东	25 965.91	西　藏	342.19
内蒙古	6091.12	河　南	15 012.46	陕　西	5465.79
辽　宁	11 023.49	湖　北	9230.68	甘　肃	2702.40
吉　林	5284.69	湖　南	9200.00	青　海	783.61
黑龙江	7065.00	广　东	31 084.40	宁　夏	889.20
上　海	12 188.85	广　西	5955.65	新　疆	3523.16
江　苏	25 741.15	海　南	1223.28		
浙　江	18 780.44	重　庆	4122.51		

时间数列数据是指在不同时间上并按时间先后顺序排列而成的数据，用以描述现象依时间而变化的情况，例如某公司 2012 年至 2017 年的产值数据(见表 2-2)。

表 2-2 某公司产值数据 单位：万元

年份	2012	2013	2014	2015	2016	2017
产值	4330.40	5023.77	6060.28	6886.31	7870.28	9353.32

平行数据则是截面数据与时间数列数据的组合，即描述多个单位在某一方面依时间而变化的情况。例如，某企业集团下各公司近几年的产值水平(见表 2-3)就是一个平行数据。

表 2-3 各公司产值的平行数据

公司	产值（万元）					
	2012	2013	2014	2015	2016	2017
A	4330.40	5023.77	6060.28	6886.31	7870.28	9353.32
B	2150.76	2578.03	3110.97	3697.62	4359.15	5050.40
C	6018.28	6921.29	8477.63	10096.11	11660.43	13709.50
D	2324.80	2855.23	3571.37	4179.52	4752.54	5733.35
E	1940.94	2388.38	3041.07	3895.55	4791.48	6091.12
F	5458.22	6002.54	6672.00	7860.85	9251.15	11023.49
G	2348.54	2662.08	3122.01	3620.27	4275.12	5284.69
H	3637.20	4057.40	4750.60	5511.50	6188.90	7065.00
…	…	…	…	…	…	…

3．观测数据与实验数据

按照统计数据的收集方法，可以将数据划分为观测数据和试验数据。其中观测数据是指通过调查和现场观测而收集的数据，一般反映社会、经济现象的数据大部分都是属于观测数据。而试验数据往往是在控制一些影响因素的情况下进行试验时所收集到的试验结果的表现。

4．初级资料和次级资料

对数据资料的收集可以从两个方面进行，一方面是收集未做过任何加工整理的原始数据资料，称为初级资料，如进行街头拦访、进行社会调查或市场调查的数据资料。初级数据资料的收集通常采用的方法有访问法、观察法、报告法等，该数据具有较强的针对性与目的性，由于其是第一手的资料，真实可信。而次级资料则是指他人为其某一特定目的而进行调查、并经过整理后的资料，或称为文案资料。次级资料的优点在于获取较为方便，成本相对较低，但缺点在于次级资料往往是为其他目的而收集的，因此在使用时要注意次级资料的适用性，包括次级资料所属的时间、性质、统计范围和统计目的等。

思考与练习

1．数据的来源有哪些？各有什么特点？

2．什么是抽样调查？ 抽样调查的主要方式方法有哪些？各有什么特点？

3．通常调查方案都包括哪些内容？

4．分析收集数据的各种方法的特点。

5．如何评价数据的质量？

6．形成调查误差的主要原因有哪些？如何提高数据的质量？

7．数据的计量尺度有哪些？各自的特点是什么？

8．数据的类型都有哪些？并举例说明各种数据类型。

9．某家电企业想通过市场调查了解新推出的新型号吸尘器投入市场后的相关情况，如该品牌的知名度、消费者的满意度、使用过程中的问题、消费者关心的问题等。为此，请你设计出一份调查方案。

10．试结合某学校或某企业的调查项目或问题，设计一份调查问卷。

第三章　数据的预处理与分组整理

第一节　统计数据的预处理

利用统计方法收集的原始数据多是零散的、不系统的，甚至可能存在有问题的数据，利用现有的二手数据也可能会因为研究目的不同、统计口径范围等问题，使我们无法直接用其分析研究对象的特征和规律。因而，在进行数据分析之前，我们还需要对所搜集的原始数据进行审核、甄别、筛选等数据的预处理工作。

一、数据的审核

数据审核就是检查数据中是否存在错误与问题，分析数据的时效性、有效性与适用性，对于不同来源的数据，数据审核的重点有所不同。

对于原始数据，主要从数据的完整性和准确性两方面进行审核。数据完整性的审核主要审核调查的个体是否有遗漏、调查的项目是否填写齐全。数据准确性审核则主要审核数据是否真实、是否有错误，审核的方法包括逻辑审核和数据平衡关系审核等。如果出现不合逻辑的异常值，则需要进行鉴别其是错误的数据还是一个正确的数据，并进行相应的处理。

对于二手数据资料，重点审核的是数据的时效性和适用性。由于二手数据往往是基于其他目的搜集并经过整理的，不一定满足我们的分析和研究目的，因而使用前需要首先了解二手数据的来源、其数据的统计范围、统计口径与统计方法等相关背景信息，不能直接简单地拿过来使用，这是适用性的审核。而时效性的审核则是指审核二手数据所属的时间，如果所获取的数据过于滞后，就可能失去使用的价值和意义。

二、数据的筛选

在数据审核中，如果发现数据错误但不能予以纠正，或者有些数据不符合调查研究但又无法弥补，就需要对数据进行筛选。通常数据筛选包括两方面的内容：一是将不符合要求的数据或有明显错误的数据剔除，如剔除考试成绩不在 0～100 范围内的异常值等；二是将符合某种特定条件的数据筛选出来，如在数据中筛选出所有女性的数据，以便对女性被调查者进行专门的统计分析与研究。

三、数据的排序

数据排序是按照一定的顺序将数据进行排序。对数据进行排序不仅可以发现数据中明显的特征和趋势，也有助于对数据的检查修正，便于对数据进行分类。在某些情况下，排序本身就是统计分析的目的，如通过统计数据的排序，找出排名前 10%的单位，表现研究对象各单位在某一方面的实力和所处的位置等。

对数据的排序有升序和降序，当然具体的排序方法需要考虑数据的具体类型，比如对于字母型的数据，习惯上升序用得比较多，因为升序与字母的自然排列相同；如果是汉字型数据，可以按汉字的首位拼音字母排列，这与字母型数据的排序完全一致，也可以就笔画多少

排序，等。对于数值型数据的排序，只有递增和递减。

利用 Excel 软件，我们可以很方便地进行数据的筛选和排序，其基本做法如下。

首先，选中数据文件中的任一单元格，单击工具栏中的【排序和筛选】按钮；然后在出现的下拉菜单中单击【筛选】或【升序】、【降序】、【自定义排序】按钮；如果进行数据筛选，则出现筛选菜单，在相应的下拉菜单中单击【全选】按钮，即去掉所有方框中的“√”，然后再单击我们要筛选出的数据，或输入筛选条件，如果在初步筛选的结果中进一步做筛选，只要重复上述步骤即可。如果仅按一个变量排序，直接单击下拉菜单中的【升序】、【降序】按钮即可；如果存在多个变量且按多个变量进行排序，则选择【自定义排序】，出现主要关键字的排序条件框，在“次序、排序依据、列”框中填入相应的条件，然后单击【添加条件】按钮，出现次要关键字的排序条件框，重复上面的操作，直到列出的条件与所选变量一致。

第二节　数据的分组

一、单一标志的分组

数据收集的目的是使研究者透过数据能够对所研究问题的内容更加清楚，并对其研究对象的本质有客观的、深入的认识。但是，实践中我们通过调查或试验搜集到的原始资料众多，且很可能是杂乱无章的，很难直观地从中看出有什么有意义的东西。特别是当数据规模比较大时，往往让人不知所措，更叫人难以消化，因而可能会导致我们不知道从哪个角度、从什么地方切入，进一步进行深入的统计分析与研究。因此，客观上需要分析研究人员对原始资料首先进行加工整理，以使其条理化、系统化，而进行此工作的重要方法之一便是统计分组。

统计分组是基于事物内在的特点和调查研究的任务与要求，根据反映研究对象各单位特征的数据水平，将研究对象划分为若干组成部分的一种统计方法。通过分组，可以将不同性质的各单位分开、相同性质的各单位归纳在一起，以反映出数据的分布特征及研究对象的基本特征，这是数据整理中极其重要的内容和方法。

统计分组的意义主要有：(1)根据数据的特点将性质不同的单位进行分组，划分为各种性质不同的类型，计算出各组中单位数在总体中所占的比重，直接反映研究对象的内部结构特征；(2)将大量数据经过分组整理后，可以更加直观、方便地表现数据总体的分布特征，反映并分析各类型、各组的数量特征；(3)可以在分组的基础上，进一步分析并揭示现象与现象之间的依存关系。

(一)分组形式

对研究对象各单位进行分组时，分组的标志只有一个的时候，则进行的是简单分组。当然，有时为了从不同侧面反映总体的特征，需要运用几个标志对数据进行分组。根据统计分组的结果，统计分组可以划分为以下几种不同的形式。

1. 简单分组

所谓简单分组，就是对研究对象中的所有单位按照一个数据标志进行分组，以反映研究对象在这一数据标志上的数据分布特征。如根据人口普查数据资料，将全部人口按年龄进行分组，可以反映出全部人口的年龄分布及其特征。

2. 平行分组体系

对研究对象中的所有单位分别按照不同的数据标志进行若干个简单分组，并将分组的结果平行排列，这样形成的一个分组体系就构成了平行分组体系。如根据人口普查数据资料，将全部人口按性别、民族、地区分别进行简单分组并平行排列就是一个平行分组体系，分组体系框架如图 3-1 所示。

3. 复合分组体系

如果将研究对象按两个或两个以上的标志层叠起来分组，称为复合分组，由复合分组形成的分组体系称为复合分组体系。在上面的例子中，先按性别分组，然后在此基础上再按地区、民族等标志进行分组即形成复合分组体系，如图 3-2 所示。

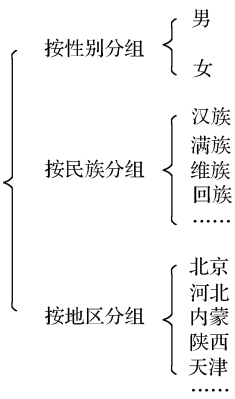


图 3-1 平行分组体系

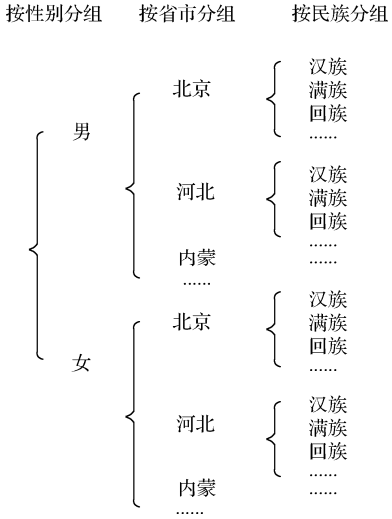


图 3-2 复合分组体系

(二)定性数据的分组列表

定性数据包括分类数据和顺序数据。定性数据本身就是对事物的一种分组，因此，只要先把所有的类别都列出来，然后统计出每一类别的数据个数(即频数)，进一步就可形成分组列表，也称为频数分布表。如将人口按性别分为男、女两组并统计出其各自的频数；按户籍所在地分为北京、河北、内蒙古、陕西、天津等。

例 3-1：我们将某校经济管理学院 2016 年新入校学生按其专业分组，并进行统计，可形成下面的频数分布表(见表 3-1)：

表 3-1 某校经济管理学院 2016 年新生按所学专业分组的频数分布表

所学专业	人数(人)
会计专业	75
经济学专业	40
经济统计专业	40
金融专业	85
财政税务专业	40
工商管理专业	30
人力资源管理专业	32
国际经济与贸易专业	30
合　　计	372

例 3-2：某街道为了解居民对社区服务质量的满意度，特在该区域抽取了 50 名居民进行了调查，让被调查者在“满意、比较满意、一般、不太满意、不满意”几个级别中进行选择。调查员随机抽到的 50 名居民对社区服务满意程度的看法如下：

满意	满意	不太满意	满意	比较满意
比较满意	比较满意	一般	不太满意	满意
比较满意	满意	比较满意	一般	满意
满意	比较满意	比较满意	一般	一般
比较满意	一般	满意	不太满意	不太满意
比较满意	满意	比较满意	一般	满意
一般	不太满意	不满意	一般	比较满意
比较满意	满意	满意	一般	不太满意
满意	不太满意	一般	一般	不太满意
满意	比较满意	满意	比较满意	一般

上面调查数据属于定性数据，已经自然形成五个组，对其汇总编制的频数分布表如表 3-2 所示。

表 3-2　满意度频数分布表

满意度	人数(人)
满意	15
比较满意	14
一般	12
不太满意	8
不满意	1
合　　计	50

通过频数分布表，我们可以了解不同类型数据的分布特征。如本例中，可以看到居民对社区服务总体是满意的，满意及比较满意的顾客共 29 人，只有 9 人表示不满意或不太满意。

对原始数据资料的整理，编制频数分布表是一种常用的方法。通过编制频数分布表，可以使统计资料得到大幅度的精简，从而使资料中蕴涵的信息能够集中概括地显示出来。

对于定性数据，我们可以用 Excel 软件直接生成频数分布表。以社区服务的满意程度数据为例，其基本作表方法如下。

首先，需要将各类别用一个数字代码来表示，如满意用“1”，比较满意用“2”，一般用“3”，不太满意用“4”，不满意用“5”表示，然后将满意程度替换成代码，并将其单独输入一列，作为接收区域。之后，在 Excel【数据分析】选项对话框中选择【直方图】命令，然后在【直方图】对话框中的【输入区域】方框内输入数据的区域，在【接收区域】方框内输入代码区域，在【输出区域】的方框内输入结果输出的位置，即可得到频数分布表。

(三)数值型数据的分组及频数分布表

前面所提到的定性数据本身往往就是对事物的一种分组，对其进行整理时，只需要进行统计计数即可。但数值型数据(包括定距数据和定比数据)在生成频数分布表前，需要先将原始数据按照某种标准划分成不同的组别，然后再统计出各组别的数据频数，形成频数分布表。

进行数值型数据的分组并编制频数分布表需要重点解决四个问题：(1)确定组数；(2)确定组距；(3)确定组限；(4)统计计数，编制频数分布表。下面结合具体的例子说明数值型数据的分组及频数分布表编制的基本要求。

例 3-3：已知有下列一组样本数据资料。

87	65	89	85	77	94	69	97
68	95	96	50	63	88	91	90
93	79	74	65	74	89	83	51
94	67	92	92	93	70	87	86
86	54	62	76	86	73	86	70
112	110	108	102	100			

要将上述数值型数据进行整理分组，编制频数分布表需要以下几步。

(1)确定组数。

对数据进行分组时，组数的确定方法主要有两种：一种是首先确定研究对象总体中各单位在所选定的分组标志下有几种质的差别，就分几组，要尽量保证组间数据资料的差异性与组内数据资料的同质性。例如，学生的考试成绩通常被认为有优、良、中、及格、不及格五种不同的质，故可分五组。另一种方法是根据数据的多少和数据的差异大小来确定分组的组数，一般数据越多、数据间差异越大，组数就越多；反之，数据越少、差异越小，则组数就相对越少。划分的组数，既不应太多，也不应太少。组数过多，达不到通过分组简化数据资料的目的；组数太少，将造成分组后资料的还原能力差，丢失的信息过多。一般情况下，组数不低于 5 组、不高于 15 组为适宜。本例中的样本数据资料，我们根据经验先确定组数为 5 或 7。

(2)确定组距。

组距，也称为组的宽度，是每组观察值的最大差，即每组观察值变化的范围：组距=组最大值–组最小值。对于分布均匀、没有特殊意义的数据，通常建议各组别的组距相同。这样，组数和组距的选择不是独立的，较多的组数意味着较小的组距；反之，较少的组数意味着较大的组距。在采取等组距分组的情况下，组距的确定方法是首先找出数据的最大值和最小值，然后利用下面的公式近似地计算出组距：

$$\text{组距}=\frac{\text{最大值}-\text{最小值}}{\text{组数}}$$

或

$$\text{组距}=\frac{\text{不小于最大值的值}-\text{不大于最小值的值}}{\text{组数}}$$

根据上式计算出来的组距可能带有小数，为了编表和计算方便，通常把它取成一个相对整的数。例如本例中，最大值是 112，最小值是 51，计算出(112–51)/7=8.71，可以以 8.71 为基础，取整数 10 作为组距。

(3)确定组限。

即确定组的上限及下限或各组的边界。组限是组与组之间的界限，或者说是每组观察值变化的范围，组限包括上组限和下组限，其中各个组的起点值为下组限，终点值为上组限。位于各组中间的点称为组中值，它可以作为各组观察值一般水平的代表，其代表性的高低，取决于组中数据变化是否呈均匀分布状态。

确定组限时一般要满足下面几点要求。①要求遵循不重复、不遗漏的原则，保证所有的数据均可以归属到其中一组，而且也只能归到一个组中。②第一组的下限应小于或等于所有数据中的最小值，最后一组上限应大于或等于所有数据中的最大值。③组限值应尽量取相对比较整的数，如 5 的整数倍、10 的整数倍等。本例中，我们确定第一组的下限为 50，最高一组的上限为 120。④若所掌握的数据会有极端值的出现，为了遵循不遗漏的原则，可以在存在极端的最小值时，将第一组设置为开口组，如 50 以下，只有上限无下限；若存在极端的最大值时，将最后一组设置为开口组，如 110 以上。⑤对于连续型变量，相邻两组的组的上下限要重叠，避免出现遗漏的现象，例中，第一组下限 50，若组距为 10，则各组分别为 50～60、60～70、70～80、80～90、90～100、100～110、110～120。

(4)在此基础上对各组数据进行统计，形成如下频数分布(见表 3-3)：

表 3-3　频数分布表

各组组别	组中值	频数	频率%
50～60	55	6	10.91
60～70	65	7	12.73
70～80	75	11	20.00
80～90	85	13	23.64
90～100	95	13	23.64
100～110	105	3	5.45
110～120	115	2	3.64
合计	—	55	100.00

频数(频率)分布表通过各组数据出现的频繁程度反映了数据的分布状况。在统计实践中，有时候仅仅有频数或频率还不够，还需要说明大于或小于某一特定值的频数或频率是多少。这时，就需要在上面的频数(或频率)分布表的基础上进一步形成累积频数(或频率)分布表。累积频数(或频率)表包括向上累积和向下累积频数(或频率)分布表，本例中的累积频率分布表如表 3-4 所示。

表 3-4　频数(频率)分布表

各组组别	组中值	频数	频率%	向上累积频率%	向下累积频率%
50～60	55	6	10.91	10.91	100.00
60～70	65	9	16.36	27.27	89.09
70～80	75	10	18.18	45.45	72.73
80～90	85	13	23.64	69.09	54.55
90～100	95	13	23.64	92.73	30.91
100～110	105	3	5.45	98.18	7.27
110～120	115	1	1.82	100.00	1.82
合　计	—	55	100.00	—	—

由上表可以看到，80 以下的数据占 45.45%，70 以上的数据占 72.73%。

对数值型数据进行分组，要能够充分显示客观现象本身存在的状态。例如，知道所研究的现象通常是服从钟形分布的，那么编制的频数分布表也应是靠近中间的频数比较多，而越往两头，组的频数应该越少。

对数值型数据，我们可以借助于 Excel 软件生成频数分布表。以表 3-3 原始数据为例，其基本作表方法如下：

首先要定义组限，确定接收区域，即单独输入一个数列，数列中每个单元格输入各组的上限，如 50～60 组，只需要输入 60；60～70 组，则输入 70 即可。其次，在 Excel【数据分析】选项对话框中选择【直方图】命令，然后在【直方图】对话框中的【输入区域】方框内输入数据的区域，在【接收区域】方框内输入各组上限所在的区域，在【输出区域】的方框内输入结果输出的位置，单击【确定】按钮即可完成。

二、两标志的交叉分组

交叉表是一种常用的分类汇总表格，最常见的是用两个标志进行分组，形成行、列方向都有分组并进行交叉而形成的分组表。将其中一个标志分组列在数据表的左侧，另一标志分组列在数据表的上部，行和列的交叉处可以对数据进行多种汇总计算，如求和、平均值、计数、最大值、最小值等。

本例中已知下面 28 人的数据资料(见图 3-3)。

	性别	家庭所在地区	平均月生活费	月平均衣物支出	买衣物首选因素
1	男	大型	800.00	200.00	价格
2	女	中小	600.00	180.00	款式
3	男	大型	1000.00	300.00	品牌
4	男	中小	400.00	40.00	价格
5	女	中小	500.00	150.00	款式
6	女	乡镇	800.00	80.00	品牌
7	男	中小	600.00	180.00	品牌
8	女	乡镇	400.00	120.00	价格
9	男	中小	1000.00	300.00	款式
10	女	大型	600.00	180.00	款式
11	女	中小	500.00	150.00	价格
12	男	乡镇	300.00	30.00	价格
13	男	乡镇	500.00	50.00	价格
14	女	中小	300.00	35.00	价格
15	男	中小	1000.00	300.00	款式
16	女	大型	800.00	350.00	款式
17	男	中小	500.00	150.00	款式
18	男	乡镇	1000.00	100.00	价格
19	女	中小	800.00	80.00	价格
20	男	乡镇	800.00	240.00	品牌
21	女	大型	500.00	50.00	品牌
22	女	大型	300.00	30.00	价格
23	男	大型	500.00	150.00	款式
24	女	中小	500.00	150.00	价格
25	男	大型	300.00	30.00	价格
26	女	大型	400.00	200.00	价格
27	男	中小	1000.00	300.00	品牌
28	男	中小	500.00	50.00	款式

图 3-3　数据资料截图

我们可以根据被调查者的性别和家庭所在地区进行交叉分组，分组结果如表 3-5 所示。

表 3-5　性别与家庭所在地区频数交叉分组表　　单位：人

计数		家庭所在地区			总计
		大型	乡镇	中小	
性别	男	4	4	7	15
	女	5	2	6	13
总计		9	6	13	28

也可在对性别和家庭所在地区分组的情况下，计算各组月平均生活费支出额，如表 3-6 所示。

表 3-6　性别与家庭所在地区分组的平均月生活费　　单位：元

项　目		家庭所在地区		
		大型	乡镇	中小
性别	男	650.00	650.00	714.29
	女	520.00	600.00	533.33

思考与练习

1．简要说明数据预处理的主要内容。
2．对数据审核的重点是什么？
3．对数据进行整理时最常用、最基本的方法是什么？
4．对数据进行分组的基本原则与要求有哪些？

第四章　数据特征的统计量描述

在对统计数据进行整理后，我们对数据分布的基本特征有了一个初步的了解，但这种了解也只是表面上的，并不能准确地描述出数据的分布。

为更深入、详细地了解数据分布的数量特征，还需要找到相应的统计量。对统计数据分布特征的描述，通常可以从三个方面进行：一是反映数据分布的集中趋势；二是反映分布的离散程度；三是反映数据分布的偏斜程度。

第一节　数据集中趋势的测度

对数据的集中趋势描述，是对总体的特征进行准确描述的重要内容之一。集中趋势是指一组数据集中于某一中心水平的倾向，测度集中趋势也是寻找数据一般水平的中心值或代表值。反映一组数据集中趋势水平的指标包括平均数、众数、中位数等。

一、平均数

平均数也称为均值，是反映数据集中趋势最常用的统计量之一，一般包括算术平均数和几何平均数两种形式。利用平均数，不仅可以反映数据集合的集中趋势、一般水平，还可以将处在不同地区、不同单位的同一现象或指标进行空间上的对比分析；可以将不同时间内的某现象进行时间上的动态对比分析，反映现象一般水平的变化趋势和规律。

(一) 算术平均数

算术平均数是将一组数据相加后除以数据的个数而得到的结果，是度量数据一般水平的常用统计量。算术平均数的计算形式包括简单算术平均数和加权算术平均数两种。

1. 简单算术平均数

若有 n 个数据 $x_1, x_2, \cdots, x_n$，则该组数据的平均数为

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

例 4-1：某单位六名职工工资水平如表 4-1 所示。

表 4-1　职工工资表

职工姓名	张舞	李荣荣	王一清	章函玉	刘民	郑志鑫
工资(元)	2860	2150	2550	3200	1800	2800

计算该单位这六名职工的平均工资：

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2860 + 2150 + 2550 + 3200 + 1800 + 2800}{6} = 2560\ (\text{元})$$

2. 加权算数平均数

当原始数据较多时，我们先对数据进行了分组，编制成了频数分布数列，这时要计算算术平均数则可以采用加权算术平均数，即将各组变量值乘以该变量值相应的频数，然后加总求和，再除以总频数。其计算公式为

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum xf}{\sum f}$$

例 4-2：有 220 名职工的工资水平如表 4-2 所示。

表 4-2　220 名职工工资表

工资(元)	1800	2150	2550	2800	2860	3200
人数(人)	30	50	80	30	20	10

计算 220 名职工的平均工资：

$$\bar{x}=\frac{\sum xf}{\sum f}=\frac{53\,870}{220}=2448.64\ (元)$$

若分组资料为组距分组资料，则我们可以取各组的组中值作为该组数据的平均水平，再代入上式计算平均数。

例 4-3：220 人的工资分组资料如表 4-3 所示。

表 4-3　职工工资分组表

工资收入(元)	人数(人)	组中值
1000～1500	30	1250
1500～2000	50	1750
2000～2500	80	2250
2500～3000	30	2750
3000～3500	20	3250
3500～4000	10	3750
合计	220	——

计算平均工资：

$$\bar{x}=\frac{\sum xf}{\sum f}=\frac{1250\times30+1750\times50+2250\times80+2750\times30+3250\times20+3750\times10}{220}$$

$$=\frac{490\,000}{22}=2227.27(元)$$

从上面的计算可以看到，加权算术平均数受两个因素的影响：一是各组数据的大小；二是各组数据出现的频数的多少。当各组数据固定不变时，则频数起着决定性的作用。出现频数较多的数据对平均数的影响作用相对大些，使平均数向其靠拢；出现频数少的数据对平均数的影响作用相对小些，平均数远离该数据。由于各水平数据出现的频数影响到了其在计算平均数时的作用，故将其称作权数。

在各组频数都相等，即 $f_1=f_2=\cdots=f_n$ 的情况下，加权算数平均数与简单算术平均数存

在着以下关系：

$$\bar{x}=\frac{\sum x_i f_i}{\sum f_i}=\frac{f_i\sum x_i}{\sum f_i}=\frac{\sum x_i}{n}$$

由此可见，简单算术平均数是加权算术平均数的一个特例，即简单算术平均数是权数相等条件下的加权算术平均数。

权数也可以用各组频数与总频数之比，即频率来表示。将各组数据乘以相应的频率后求和，即得到加权算术平均数，其计算公式为

$$\bar{x}=\sum x_i\cdot\frac{f_i}{\sum f_i}$$

在频数分布数据中，各组数据出现的频数与总频数同时发生变化，而频率不变时，则加权算术平均数的数值不变。

权数是计算算术平均数的核心问题，权数体现了各组变量值在数据集合中的重要程度。计算算术平均数的权数主要有两类：客观权数和主观权数。

客观权数是指与被平均的数据存在客观联系的指标。客观权数的确定可以从以下两个方面考虑：一是在频数分布中，以各组数据出现的频数或频率为权数；二是根据事物之间的相互联系，选择与被平均的数据存在直接数量关系的指标为权数。如计算某公司所属的几个分公司的平均利润率，资料如表 4-4 所示。

表 4-4　公司利润表

利润率(%)	分公司数(个)	职工人数(人)	销售额(万元)
5 以下	2	200	300
5～10	8	500	5000
10～15	9	600	8000
15 以上	1	120	400
合计	20	1420	13 700

平均利润率的含义为所有分公司单位销售额所获取的利润。利润率等于利润额除以销售额。利润率与销售额存在直接数量关系，而与分公司数和职工人数没有直接关系。计算利润率的平均数并不改变利润率本身的意义，因此应该选用销售额为权数，而不用分公司数或职工人数为权数。该公司的平均利润率：

$$\bar{x}=\frac{2.5\%\times300+7.5\%\times5000+12.5\%\times8000+17.5\%\times400}{13\,700}=10.6\%$$

在有些情况下，缺少或不存在客观权数的资料，但又需要体现被平均对象在总体中的重要程度，这时则需要根据人们的经验设定权数。这种权数称为主观权数。一般来讲，在总体中作用较大的赋予较大的权数，作用较小的赋予较小的权数。

例如，综合评价学校的教学工作情况，可以将教学工作分解为教学条件、教学状态和教学效果三个方面。采取评分的方法，各项内容满分为 100，每项内容赋予不同权数，教学条件和教学状态权数设为 3，教学效果的地位显得更重要些，权数设为 4。然后根据各项内容所评的分数进行加权平均得出平均分，来综合评价学校的教学工作情况。若某校教学工作三

个方面的得分是：教学条件 80 分，教学状态 85 分，教学效果 90 分，则教学工作综合得分：

$$\bar{x}=\frac{80\times3+85\times3+90\times4}{3+3+4}=85.5$$

算术平均数是根据全部数据计算得到的，因此受数据中极端数值的影响较大。当数据中存在极大值时，会使计算出的算术平均数偏大；当数据中存在极小值时，会使计算出的算术平均数偏小。无论出现极大值还是极小值，计算的平均数都不能正确反映数据集合的一般水平，也就不能准确测定数据集合的集中趋势。因此，在实际运用算术平均数时，如果存在过大或过小的数据，或将其剔除，然后计算余下的数据的平均数，并称其为切尾平均数；或采用其他集中趋势的度量指标，如下面要介绍的众数或中位数等。

算术平均数是将所有数据进行平均，所以其适用于定量数据计算平均数，而定类数据和定序数据无法通过计算算术平均数来说明总体的集中趋势。

（二）几何平均数

几何平均数是 n 个数据乘积的 n 次方根。几何平均数有两种：简单几何平均数和加权几何平均数。

1. 简单几何平均数

若数据集合中每个数据只出现一次，计算其几何平均数应采用简单几何平均法，其计算公式为

$$G=\sqrt[n]{x_1\cdot x_2\cdot\cdots\cdot x_n}=\sqrt[n]{\prod x_i}$$

式中，G 表示几何平均数；π 表示连乘积符号。

例 4-4：某企业生产某产品的整个生产过程包括连续流水作业的三道工序，这三道工序分别在三个车间加工，三个车间的产品合格率分别为 80%、95%和 90%。则这三个车间的平均合格率为

$$G=\sqrt[3]{0.8\cdot0.95\cdot0.9}\times100\%=88.1\%$$

例 4-5：某企业 2011—2015 年产值的发展速度分别为 116%、110%、115%、112%、108%，计算该企业产值的年平均发展速度：

$$G=\sqrt[5]{116\%\cdot110\%\cdot115\%\cdot112\%\cdot108\%}=112.16\%$$

2. 加权几何平均数

当数据较多且每个数据出现的次数不同时，计算平均数应采用加权几何平均法。其计算公式为

$$G=\sqrt[f_1+f_2+\cdots+f_n]{x_1^{f_1}\cdot x_2^{f_2}\cdot\cdots\cdot x_n^{f_n}}=\sqrt[\sum f_i]{\prod x_i^{f_i}}$$

例 4-6：某银行在 10 年内几次调整贷款利率(按复利计息)，其各年的利率：第一年 8%，第 2 年至第 5 年为 6.5%，第 6 年至第 8 年为 5%，第 9 年至第 10 年为 4%，这 10 年的平均贷款利率的计算如下：

$$G=\sqrt[1+4+3+2]{1.08^1\cdot1.06^4\cdot1.05^3\cdot1.04^2}=1.0569$$

平均贷款利率为：　　$105.69\%-100=5.69\%$

几何平均法是用各变量值乘积开方的方法计算平均数的，因此，当总体中有一变量值为零时，几何平均数的计算结果则为零。在社会经济问题的研究中，这种计算无意义。

几何平均数与算术平均数的应用条件不同。算术平均数应用于按算术级数形式变化的事物，即事物总量等于各变量值的总和，求平均水平。而几何平均数应用于按几何级数形式变化的数据，即事物总量等于各变量值的乘积。在社会经济领域，几何平均数常用来计算平均比率或平均速度。

二、中位数

中位数是指将数据集合中的所有数据按大小顺序排列后，处于中点位置上的那个数据。中位数将数据分为相等的两部分：一部分数据小于中位数；另一部分数据大于中位数。确定中位数的方法如下。

1. 根据未分组的原始数据确定中位数

首先将数据按大小顺序排列，确定中位数所在的位置，然后根据中位数的位置找出对应的数据即中位数。

当数据的个数 n 为奇数时，中位数所在的位置为 $\frac{n+1}{2}$，该位置上的数值为中位数；当数据的个数 n 为偶数时，中位数的位置有两个，一个是 $\frac{n}{2}$，另一个是 $\frac{n}{2}+1$，这两个位置上的数值的平均数则为中位数。

例如：某单位 6 个职工的工资水平如表 4-5 所示。

表 4-5 职工工资表

职工姓名	张舞	李荣荣	王一清	章函玉	刘民	郑志鑫
工资收入	2860	2150	2550	3200	1800	2800

将职工工资按大小顺序排列：1800、2150、2550、2800、2860、3200，中间位置为 6/2=3 和 6/2+1=4，则中位数为(2550+2800)/2=2675 元。

2. 根据单项式频数分布数列确定中位数

例如，已知 323 个家庭的人口数资料编制成单项式频数分布数列如表 4-6 所示。

表 4-6 家庭人口数分布表

家庭人口数	家庭数	累计数
1	25	25
2	75	100
3	180	280
4	33	313
5	10	323
合计	323	—

对于这样的单项式分布数列的数据计算其中位数，首先确定中位数所在的位置 $\frac{n+1}{2}=(323+1)/2=162$，然后在数列中找到中位数所在的具体位置上的值，即计算累计频数为 162 的组所对应的数据，本例中为第三组的数据 3。故 323 个家庭人口数的中位数为 3 人。

3．根据组距分组数列确定中位数

若按年龄将 1300 人分为五组，形成组距数列如表 4-7 所示。

表 4-7　年龄分布表

年龄	人数	向上累计	向下累计
20～30	120	120	1300
30～40	400	520	1180
40～50	530	1050	780
50～60	200	1250	250
60 以上	50	1300	50
合计	1300	—	—

要计算这 1300 人的平均年龄，由于无法知道其具体的年龄数据，因而只能借助于分组数据近似地估计其中位数。其计算方法如下。

首先确定中位数所在组，计算公式是 $\frac{\sum f}{2}$，然后计算中位数的近似值，计算时可以采用下限公式，也可采用上限公式。

$$\text{下限公式：}\ M_e = L + \frac{\frac{\sum f}{2} - S_{m-1}}{f_m} \cdot d$$

$$\text{上限公式：}\ M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} \cdot d$$

式中，L 表示中位数所在组的下限；

U 表示中位数所在组的上限；

f_m 表示中位数所在组的频数；

$\sum f$ 表示各组频数之和；

S_{m-1}、S_{m+1} 分别表示中位数所在组以前、以后各组的累计频数；

d 表示中位数所在组的组距。

本例中，我们利用上限公式，计算中位数如下。

首先确定中位数所在的位置：$\frac{\sum f}{2} = \frac{1300}{2} = 650$，故中位数所在组为第三组，即 40～50 岁这一组。然后利用上限公式计算中位数的估计值：

$$M_e = U - \frac{\frac{\sum f}{2} - S_{m+1}}{f_m} \cdot d = 50 - \frac{\frac{1300}{2} - 250}{530} \times (50 - 40) = 42.45\ (\text{岁})$$

由于中位数是根据所有数据中点位置确定的，因而，中位数不受极端数值的影响。当总体分布偏斜程度较大时，中位数对于测定所有数据的集中趋势，反映数据集合的一般水平具有较大的实用性，即在有极端数值出现时，中位数作为分析现象中集中趋势的数值，比平均

数更具有稳健性。此外，中位数是根据数据大小顺序排列后确定的，因此其适用于定序数据及定量数据，而定类数据则不能使用中位数。

三、众数

众数是数据集合中出现次数最多的数据。根据所掌握的数据资料的性质不同，众数有不同的计算方法。

(1)根据单项式变量数列确定众数。确定单项式变量数列的众数比较简单，只需要找出次数最多的数据即可。

(2)根据组距分布数列估计众数。组距分布数列确定估计众数的近似值较单项式数列确定众数要复杂一些。估计的具体方法是，首先根据数列中各组频数确定众数所在的组，然后利用公式计算出众数的近似值。其计算公式如下。

$$\text{下限公式：}\ M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot d$$

$$\text{上限公式：}\ M_0 = U - \frac{\Delta_2}{\Delta_1 + \Delta_2} \cdot d$$

式中，L 表示众数所在组的下限；

U 表示众数所在组的上限；

Δ_1、Δ_2 分别表示众数所在组的频数与前一组、后一组频数之差；

d 表示众数所在组的组距。

上例中，40～50 岁的组有 530 人，最多，则该组为众数所在的组，将有关数据代入下限(或上限)公式中计算众数。

$$\text{下限公式：}\ M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \cdot d = 40 + \frac{530 - 400}{(530 - 400) + (530 - 200)} \times (50 - 40) = 42.83\ (\text{岁})$$

由于众数容易求得，一般采用直接观测法即可，它不仅适合于定量变量，也适合反映定类数据的集中趋势。众数是从数据出现次数的多寡上反映现象的集中趋势的，因此，当分布数列中有一组数据具有非常大的频数时，往往优先采用众数反映其集中趋势。

众数是根据数据出现的频数的多少确定的，因此不受极端值的影响，在组距数列中，各组分布的频数受组距大小的影响，所以根据组距分布数列估计众数时，要保证各组组距必须相等。若出现多个众数，为了认识事物的本质特征，可将其分解为两个或几个不同的分布加以分析与研究。

四、算术平均数、中位数和众数的关系

平均数、中位数和众数各自具有不同的特点，如平均数包含了样本的很多信息，但它容易受少数极端值的影响。中位数是数据按照大小排列之后位于中间的那个数，它不易受极端值影响，所以说中位数比平均数稳健。除了中位数和平均数之外，反映数据集中趋势的还有众数，众数反映的信息不多，有时也不一定存在唯一众数，尤其是连续变量，可能没有重复的数据，这时也不可能有众数。实践中众数用得不如平均数和中位数普遍，但对于定性变量，由于数据的性质，只能计算频数或频率，因此通常使用众数。

使用哪种统计量或测度指标来反映数据的集中趋势，要根据我们所掌握的数据的类型来确定，如对于分类数据，只适用于计算众数，而不能计算中位数和平均数。对于顺序数据，只适用于计算众数和中位数，不能计算平均数。对于数值型数据，主要是用平均数作为集中趋势的测度值，此外也可以利用众数、中位数等。

从分布的角度看，众数始终是一组数据分布的最高值，中位数是处于一组数据中间位置上的值，而均值则是全部数据的算术平均。三者之间的关系：对于具有对称单峰分布特征的数据集合，这三个统计量应该大体上差不多。而如果单峰的分布形状在右边拖尾，那么一般来说，中位数小于平均数。反过来，如果在左边拖尾，则一般平均数小于中位数。也就是说，和中位数相比，平均数总是偏向长尾巴那边。三个统计量测度指标在不同分布状况下的关系如图 4-1 所示。

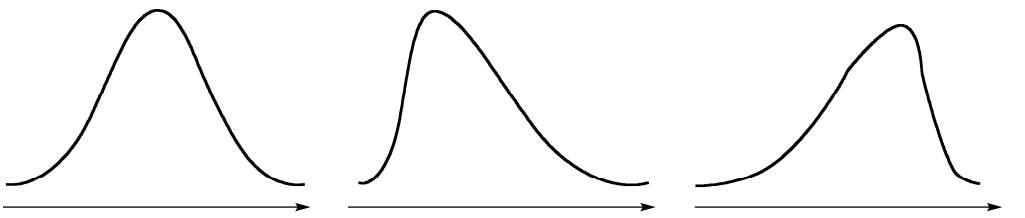


图 4-1　分布图

第二节　数据离散程度的测度

集中趋势的统计量反映了数据聚集的中心位置与数值。数据的离散程度是数据分布的另一重要特征，它所反映的是各数据值远离其中心值的程度，即数据的差异程度。数据的离散程度大，集中趋势的测度值对该组数据的代表性就越差，离散程度越小，其代表性就越好。要反映数据集合中各数据之间的差异状况，我们需要通过计算关于数据离散程度的统计量。描述数据离散程度的统计量主要有极差、四分位差、方差、标准差等。

一、极差与四分位差

1. 极差

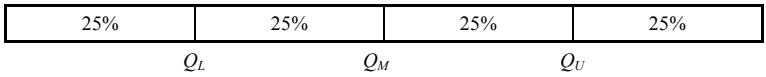
极差是数据集合中的数据最大值与最小值之差，即 $R = x_{\max} - x_{\min}$ 。

极差说明了总体中所有数据值的变动范围。极差越大，说明数据的变动范围越大，从而说明数据的差异越大，反之则越小。

极差的计算简单，但它是由全部数据中的两个极端值决定的，并未考虑到其他数据的差异。当两个极端值差异较大，而其他数据却集中于较小区间内或均匀分布在两个极端数值之间时，极差则不能确切地反映全部数据之间的差异。

2. 四分位差

四分位差是上四分位数(即一组数据中 75%位置上的数据)与下四分位数(数据中 25%位置上的数据)之差，也称为内距。其中，四分位数是在数据按大小顺序排列后，将全部数据等分为四部分的三个点上的数据。显然，中间的四分位数就是中位数，因而通常所说的四分位数是指处在 25%位置上的数值和处在 75%位置上的数值。



四分位数与中位数的计算方法类似，根据原始数据计算四分位数时，首先对数据进行排序，然后确定四分位数所在的位置，该位置上的数据就是四分位数。一般，25%和 75%位置上的四分位数位置的确定方法如下。

下四分位数(25%分位数)位置：

$$Q_L\text{位置}=\frac{n}{4}$$

上四分位数(75%分位数)位置：

$$Q_u\text{位置}=\frac{3n}{4}$$

如果位置是整数，四分位数就是该位置对应的数值，如果是在整数加 0.5 的位置，则取该位置两侧值的平均数；如果是在整数加 0.25 或 0.75 的位置，则四分位数等于该位置前面的数据加上按比例分摊位置两侧数值的差值。

因为四分位数计算的是中间 50%数据的极差，因此与极差相比，它基本不受极端值的影响，但是与极差一样，四分位极差也仅仅是通过两个数据之差反映数据的离散程度，没有反映全部数据或大部分数据之间的差异程度。

二、方差与标准差

方差和标准差是反映数据离散程度最常用、最重要的测度统计量或指标。方差是一组数据中各数值与其算术平均数离差平方的平均数，标准差是方差的平方根。根据总体数据计算的称为总体方差或总体标准差；根据样本数据计算的称为样本方差或样本标准差。用 σ^2 表示总体方差，用 σ 表示总体标准差，s^2 代表样本方差，s 代表样本标准差。

(一)总体方差和标准差

计算总体方差和标准差时，根据所拥有的数据资料不同，可以选择不同的计算方法。方差的计算方法有两种：简单平均法和加权平均法。

1. 简单平均法

根据未分组的原始数据计算总体方差，是计算每个数据与算数平均数的离差平方的简单算数平均数。其计算公式：

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}=\frac{\sum x_i^2}{n}-\left(\frac{\sum x_i}{n}\right)^2$$

式中，σ^2 表示方差；n 表示数据个数。

与方差不同，标准差与变量值的计量单位相同，其实际意义比方差清楚。因此，在对社会经济现象进行分析时，我们更多地使用标准差：

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

例 4-7：有六名学生的统计学考试成绩分别为 68，72，78，84，88，90，该六名学生成绩的方差和标准差计算如下。

首先，计算平均成绩：　　$\bar{x}=\frac{\sum x_i}{n}=80$

然后按方差的计算公式计算方差：

$$\begin{aligned}\sigma^2&=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\\&=\frac{(68-80)^2+(72-80)^2+(78-80)^2+(84-80)^2+(88-80)^2+(90-80)^2}{6}\\&=65.33\\\sigma&=\sqrt{65.33}=8.08\end{aligned}$$

若同时，该六名同学的数学平均成绩也是 80，但数学成绩的标准差为 10，则说明统计学成绩的差异程度小于数学。

2. 加权平均法

根据分组后的数据频数分布数列计算方差，需要采用加权平均法。其计算公式为

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2f_i}{\sum_{i=1}^{n}f_i}=\frac{\sum x_i^2f_i}{\sum f_i}-\left(\frac{\sum x_if_i}{\sum f_i}\right)^2$$

表 4-8 是某班 50 名学生的统计学成绩的分组资料。

表 4-8　成绩分布表

分数	人数 f	组中值 x
60 以下	2	55
60～70	8	65
70～80	22	75
80～90	14	85
90～100	4	95
合计	50	

计算该班学生成绩的方差过程如下。

首先，计算平均成绩：

$$\bar{x}=\frac{\sum x_if_i}{\sum f_i}=77(\text{分})$$

然后，利用方差的计算公式得到：

$$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{n} f_i}=\frac{4400}{50}=88$$

标准差 $\sigma=\sqrt{88}=9.38$ 。

（二）样本方差与标准差

样本的方差和标准差与总体的方差和标准差的计算是有区别的，总体方差是用数据个数或总频数去除离差平方和，而样本方差通常是用于估计或推断总体方差，为了得到总体方差的无偏估计量，样本方差的计算是用样本数据个数或总频数减 1 去除离差平方和，其中样本数据个数减 1 即 n–1 称为自由度。这样，样本方差及标准差的计算公式如下。

未分组数据的方差：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}$$

未分组数据的标准差：

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}=\sqrt{\frac{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n-1}}$$

分组数据的方差：

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{n} f_i-1}$$

分组数据的标准差：

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2 f_i}{\sum_{i=1}^{n} f_i-1}}$$

若上例数据是从全校学生中随机抽取出的 50 名学生的统计学成绩数据，目的是用于推断全校学生的统计学的成绩，这时计算样本方差和标准差需要采用样本数据计算方差和标准差的计算公式。

三、离散系数

标准差是反映数据离散程度的绝对值，其数值的大小不仅受数据本身差异大小的影响，还受到其他两方面因素的影响。一方面受数据本身水平高低的影响，即与变量值的均值大小有关。数据总体绝对水平高的，其离散程度的测定值一般也就大，绝对水平小的数据集合，其离散程度的测定值一般也就小；另一方面，它们与数据的计量单位有关，采用不同计量单位的数据，其数据离散程度的测定值也就不同。

当比较几组数据的离散程度大小时，需要消除变量值水平和计量单位对离散程度测定值的影响，计算离散程度的相对指标，即离散系数。

离散系数是测定数据集合中数据离散程度的相对指标，可以用数据的标准差与其相应的平均数之比来表示。其计算公式：$V_\sigma = \frac{\sigma}{\overline{x}}$。

离散系数与平均数代表性的优劣呈反方向关系，离散系数大，说明数据内部的差异程度大，数据的稳定性差，平均数的代表性差；离散系数小，说明数据内部的差异程度小，数据的稳定性强，平均数的代表性强。

例如，甲、乙两个地区职工的月平均工资分别是 3200 元和 3500 元，其职工工资的标准差分别为 210 元和 220 元，要分析哪个地区职工工资的差异大，需要计算离散系数。

$$V_\sigma(\text{甲}) = \frac{\sigma}{\overline{x}} = \frac{210}{3200} = 0.065625$$

$$V_\sigma(\text{乙}) = \frac{\sigma}{\overline{x}} = \frac{220}{3500} = 0.062857$$

经过比较，可以看到，乙地区的平均工资高于甲地区，且乙地区职工工资的差异程度要小于甲地区。

第三节　数据分布形状的度量

进一步地，还需要了解、描述数据分布是否对称，如果分布不对称，其偏斜程度如何，其分布的扁平程度又如何，等等。要做这样的分析，就需要计算相应的统计量，包括偏态系数和峰度系数。

一、偏态系数

偏态系数是用来反映数据分布偏斜程度的统计量。当数据是单峰分布时，有对称分布和非对称分布两种，其中非对称分布称为偏态分布，包括右偏(正)分布(即右拖尾)和左偏(负)分布(即左拖尾)，如图 4-2 所示。

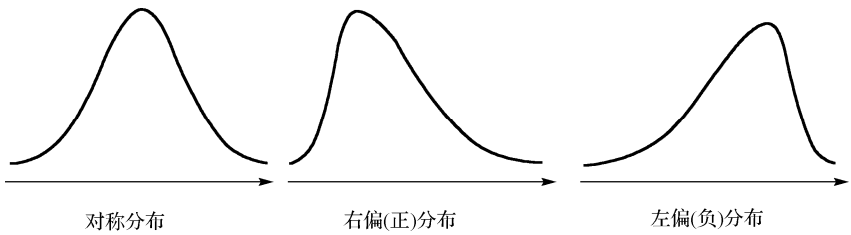


图 4-2　分布图

在本章第一节数据集中趋势的测度一节中已经提到，数据呈现对称分布时一个基本的特征是平均数、中位数和众数相等；如果数据呈现右偏分布，则其众数在左边，小于中位数和平均数，而平均数一般大于众数和中位数；如果数据呈现左偏分布，则其众数在右边，往往大于平均数和中位数，而平均数通常小于中位数和众数。

为了准确地测定分布的偏斜程度和进行比较分析，需要计算偏态系数，常用的统计量如下。

1. Pearson 偏度系数

Pearson 偏度系数是以标准差为度量单位计算的众数与算数平均数的离差，其计算公式：

$$\mathrm{SK}=\frac{\bar{x}-M_0}{\sigma}$$

SK 通常取值在–3～+3 之间，其绝对值越大，表明偏斜程度越大，反之表明偏斜程度越小。

当 SK=0 时，分布为对称分布；SK<0 时，分布呈左偏分布，或称负偏态；SK>0 时，分布呈右偏斜分布，或称为正偏态。

2. 矩法偏度系数

Pearson 偏度系数的思想比较容易理解，但精确程度不高。矩法偏度计算方法能够弥补这一不足，其计算公式：

$$\mathrm{SK}=\frac{\sum(x_i-\bar{x})^3 f_i}{\sum f_i}\div S^3$$

式中，$\dfrac{\sum(x_i-\bar{x})^3 f_i}{\sum f_i}$ 为三阶中心矩，

S^3 为标准差的三次方。

可以看到，矩法偏度就是三阶矩与标准差的三次方之比。

同样，当 SK=0 时，分布为对称分布；SK<0 时，分布呈左偏分布，或称负偏态；SK>0 时，分布呈右偏斜分布，或称为正偏态。

二、峰度系数

在社会经济现象中，许多变量数列的分布曲线与正态分布曲线相比，其顶部的形态会有所不同，而这种差异通常具有重要的社会经济意义。

在本章前面两节中我们已经了解到，集中趋势越明显，离散程度越小，频数分布的形状就越高耸；而集中趋势越不明显，离散程度越大，反映出来的频数分布就显得越加扁平。峰度系数可以反映数据分布峰值的高低，可以用来说明数据分布曲线的顶端尖削或扁平程度。以标准正态分布为参照标准，比正态分布尖削的分布为尖峰分布，比正态分布扁平的分布为平顶分布(如图 4-3 所示)。

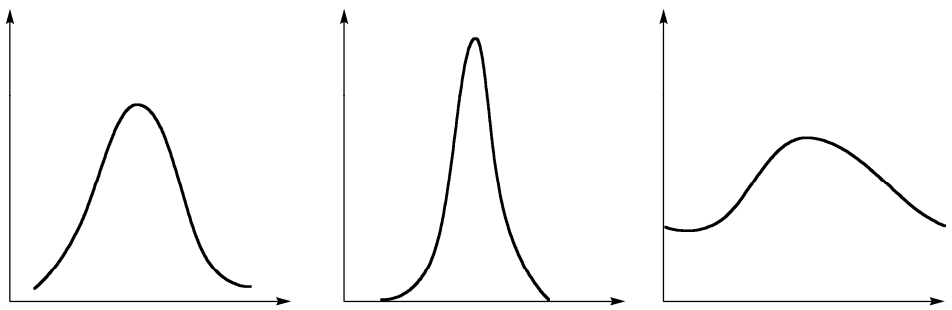


图 4-3　分布图

峰度的测量指标，常常可用标准差的四次方除以四阶中心矩的方法来计算，计算公式：

$$\beta = \frac{m_4}{\sigma^4}$$

式中，$m_4 = \dfrac{\sum (x_i - \bar{x})^4 f_i}{\sum f_i}$，$\sigma = \sqrt{\dfrac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$。

一般，当 β=3 时，数据的分布峰度表现与正态分布相同；当 β>3 时，为尖顶分布，表明数据分布曲线的顶部较正态分布曲线更为陡峭，且 β 越大，顶部就越陡峭；当 β<3 时，为平顶分布，表明数据分布在众数附近比较分散，使得频数分布曲线的峰顶较正态分布曲线平缓，且 β 值越小，顶部就越加平坦。

第四节　描述数据特征的统计量的计算与应用

本章一、二、三节介绍了描述数据特征的统计量及计算方法与公式。在计算中，当数据较少时我们可以利用计算器计算这些指标，而当数据较多时我们可以直接采用相关的软件实现计算。

一、用 Excel 计算

例如，有如下 X、Y 两组数据列在 Excel 数据表中(如图 4-4 所示)。

	A	B	C	D	E	F	G	H	I
1	x	234	143	187	161	150	228	153	16
2	y	159	198	160	152	161	162	163	19

图 4-4　EXCEL 数据表截图 1

在数据分析工具的菜单中，选择【描述统计】按钮(如图 4-5 所示)。

确定数据区域与输出区域后，单击【确定】按钮即可得到计算结果(如表 4-9 所示)。

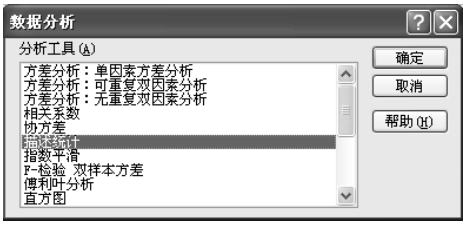


图 4-5　Excel 操作截图 2

表 4-9　计算输出结果表

X		Y	
平均	181.83	平均	177.00
标准误差	5.87	标准误差	4.63
中位数	178.50	中位数	168.00
众数	187.00	众数	165.00
标准差	28.78	标准差	22.69
方差	828.32	方差	515.04
峰度	-1.11	峰度	0.72
偏度	0.34	偏度	1.33
区域	93.00	区域	81.00
最小值	141.00	最小值	152.00
最大值	234.00	最大值	233.00
求和	4364.00	求和	4248.00
观测数	24.00	观测数	24.00

计算结果显示出了两个变量的平均数、中位数、众数、标准差、方差、偏度系数、峰度系数等描述指标，通过这些指标我们可以对两组数据的特征进行描述、对比和分析。

二、用 SPSS 软件计算

(1)打开或输入数据后，在 Analyze 下拉菜单中，选择描述统计【Descriptive Statistics】，再选择【Descriptives】选项(如图 4-6 所示)。

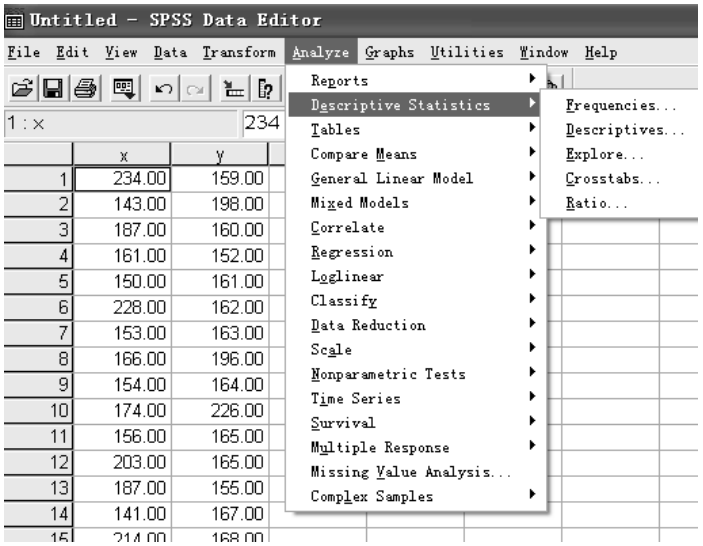


图 4-6 SPSS 操作截图 1

(2)将要描述的数据变量单击选入【Variable】中(如图 4-7 所示)。

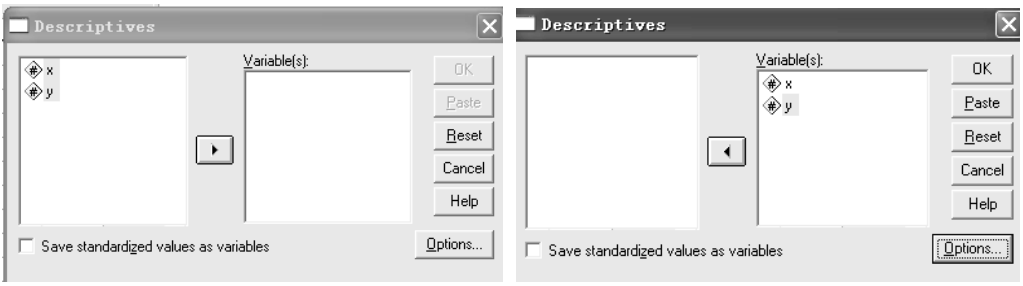


图 4-7 SPSS 操作截图 2

单击【OK】按钮，即可得到主要的数据描述统计量(如图 4-8 所示)。

Descriptive Statistics

	N	Range	Mean	Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
x	24	93.00	181.8333	28.78053	828.319	.336	.472	-1.107	.918
y	24	81.00	177.0000	22.69457	515.043	1.333	.472	.718	.918
Valid N (listwise)	24								

图 4-8 SPSS 操作截图 3

思考与练习

1．描述数据集中趋势的统计量都有哪些？
2．举例说明平均数、中位数和众数的特点及应用条件。
3．描述数据离散程度的统计量都有哪些？
4．举例说明极差、四分位差、方差与标准差的特点及应用条件。
5．简要说明为什么要计算离散系数。

6．如何理解权数的意义？在什么情况下，应用简单算数平均法和加权算数平均法计算的结果是一样的？如何确定权数？

7．已知下面的统计资料，试进行统计数据的整理，并利用合适的统计量进行描述性的统计分析。

序号	性别	统计学课成绩	论文成绩	序号	性别	统计学课成绩	论文成绩
1	男	92	优秀	18	男	74	中等
2	男	75	良好	19	男	92	优秀
3	女	62	合格	20	女	90	中等
4	女	90	中等	21	男	76	中等
5	女	62	合格	22	男	62	合格
6	女	84	良好	23	女	63	中等
7	女	92	优秀	24	女	77	中等
8	男	84	中等	25	女	81	优秀
9	男	74	良好	26	女	88	中等
10	男	86	优秀	27	女	92	优秀
11	男	71	合格	28	男	82	中等
12	女	95	良好	29	男	86	良好
13	女	72	中等	30	男	71	中等
14	女	84	中等	31	女	83	优秀
15	女	81	优秀	32	女	83	中等
16	女	75	合格	33	女	71	良好
17	女	76	良好	34	女	88	良好
				35	女	93	良好

8．某企业某班组工人日产量资料如下表所示：

按日产量分组	工人数
50～60	5
60～70	13
70～80	19
80～90	10
90～100	8
合计	55

根据上面的统计数据资料，指出：

(1)表中的变量数列属于哪种变量数列？

(2)表中的变量、变量值，各组的上限、下限、组中值。

(3)计算组距、频率、向上累积频率、向下累积频率。

(4)计算日产量的平均数和标准差。

第五章　数据资料的图形显示

对原始数据资料进行加工整理，可以使其条理化和系统化，如果在此基础上进一步地用一个精简的、直观的方法——统计图来显示出数据所表现的重要内容和特征，则具有更重要的意义和作用。当然，若描述数据特征时采用了不恰当的或不适宜的方法，极有可能会使事实受到扭曲，形成统计陷阱。所以说，应该如何对数据进行综合、描述和展示，是统计工作中要解决的重要问题之一。

第一节　定性数据的统计图示

对数据进行描述不仅可以用统计指标和统计表，还可以用统计图的方法。在统计实践中，用一张好的、适合的统计图显示所收集到的数据特征有时比直接用统计数据或数据的频数分布表等手段会更加形象和直观，甚至可能会得到意想不到的效果。统计图很多，但各有各的使用条件和适用对象，一般常用的、适用于定性数据(分类数据和定序数据)的统计图主要有条形图、柱形图、帕累托图、饼图、环形图等。

一、条形图与柱形图

条形图与柱形图是用宽度相同的条形的长度表示各类数据多少或各指标大小的一种统计图形，主要用于观察数据的分布或进行各项信息的比较等。

在绘制条形图与柱形图时，各类别(或指标)名称可以放在纵轴，也可以放在横轴上，没有尺度，只用来表示各类数据的名称和各项指标信息的名称。各类别名称放在横轴时通常被称为柱形图，放在纵轴时称为条形图。而相应各类数据的多少、指标的大小则用图中长方形的长度(或高度)来表示。

一般，反映定性数据分布特征时多用柱形图，而用量观察各项信息大小时则采用条形图。而如果数据是对同一事物在若干时间点或段上的度量，则一般以横坐标表示时间，纵坐标表示数据的大小，即应使用柱形图。

例 5-1：2016 年某集团公司下属四个分公司的职工人数如表 5-1 所示。

表 5-1　职工人数表

分公司	甲	乙	丙	丁
人数(人)	263	420	580	380

表中数据用柱形图表示为如图 5-1 所示。

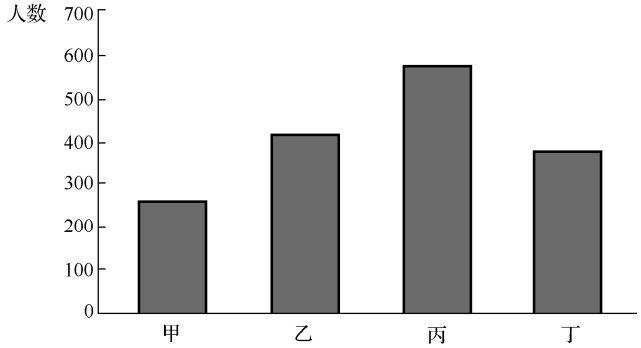


图 5-1　分公司职工人数分布的柱形图

用条形图显示上述数据的结果如图 5-2 所示。

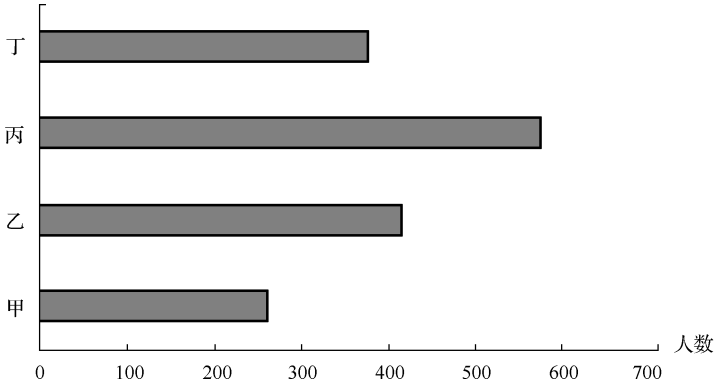


图 5-2　分公司职工人数分布的条形图

二、帕累托图

帕累托图是按各类别数据出现的频数的多少排序后绘制的柱形图。通过对各条形的排序，容易看出哪类数据出现得多，哪类数据出现得少，如图 5-3 所示。

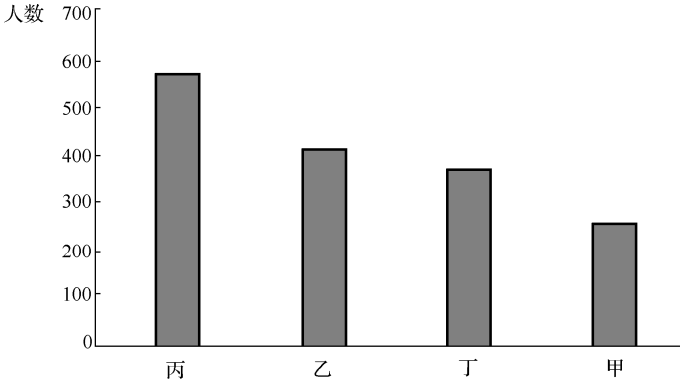


图 5-3　分公司职工人数分布的帕累托图

三、饼图

饼图一般是用来描述和表现研究对象的内部构成的，即各类数据占全部的百分比。如满意人数占被调查者总人数的百分比、各分公司职工人数占总公司全部职工人数的百分比、同

一种产品各品牌的市场占有率等。饼图是用圆形及圆内扇形的角度来表示数据集合内部构成的图形，对于反映、研究结构性问题十分有用。如居民对社区服务满意度情况构成的饼图，如图 5-4 所示。

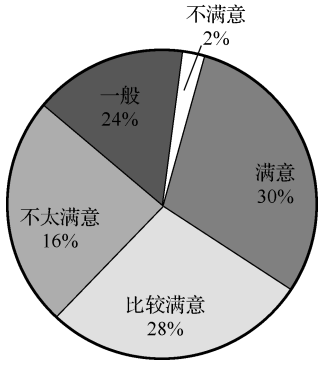


图 5-4 居民对社区服务满意度的饼图

使用饼图时必须注意：

(1)饼图中的分类数量最好不要太多，若分类数目过多时，可以采取的做法是从这些分类中选择其中几个最重要的，然后把剩余的部分全部合并成一类并称为“其他”；

(2)各类所占百分比之和必须是 100%；

(3)各类型数据所占百分比需要与扇形区域面积的比例基本一致。

四、环形图

饼图只能显示数据集合中各类数据所占的百分比，比如，要反映 A 地区居民对社区服务满意度不同态度的构成状况，可以用饼图来表示，但若要比较 A 地区、B 地区、C 地区、D 地区四个地区的居民对社区服务满意度态度的构成差异，则需要绘制 4 个饼图，但这种做法既不经济也不便于比较。为便于比较，我们可以把饼图叠在一起，挖去中间的部分，形成环形图。环形图与饼图类似，但又有区别。每个地区的数据集合用一个环来表示，一个地区的每一类数据用环中的一段表示。因此环形图可显示多个数据集合各部分所占的相应的百分比，从而有利于各数据集合构成的比较。

例 5-2：表 5-2 是 A、B、C、D 四个地区居民对社区服务的满意度构成的统计数据资料，绘制环形图比较四个地区的差异。

表 5-2 各地区对社区服务满意度几种态度的构成(%)

满意度	地区			
	A	B	C	D
满意	30	28	18	15
比较满意	28	25	25	30
一般	24	24	26	24
不太满意	16	13	18	15
不满意	2	10	13	16

根据上面的数据绘制环形图如图 5-5 所示，图中从里向外的四个环分别是 A、B、C、D 四个地区的满意度构成。

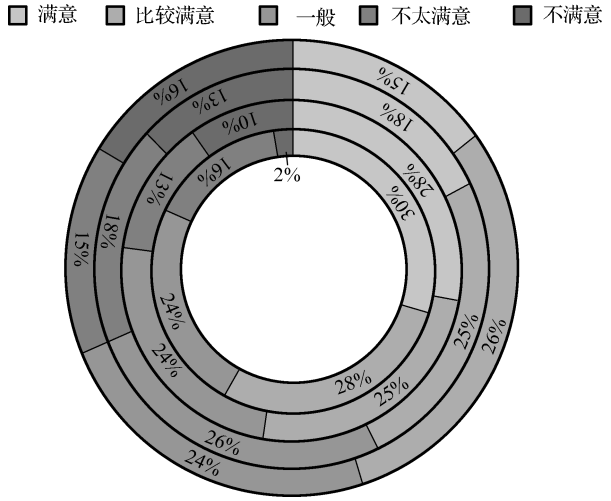


图 5-5　四个地区居民对社区服务满意程度构成环形图

第二节　数值型数据的统计图示

上节中介绍的定性数据的图示方法，也都适合于数值数据，但数值型数据还有一些特定的图示方法，它们并不适用于定性数据。

一、直方图

定性变量的取值相对较少，因而我们通常可以用饼图、柱图或条形图来表示数据的分布，但是像收入、产值、身高等这些定量变量的可能取值太多，因而这类数值型数据的分布可以考虑直方图。

直方图是用于展示数值型数据分布的一种常用图形，它是用矩形的宽度表示数据的组，用高度来表示频数的数据分布图。通过直方图可以观察数据分布的大体形状，如分布是否对称等，如图 5-6 所示。

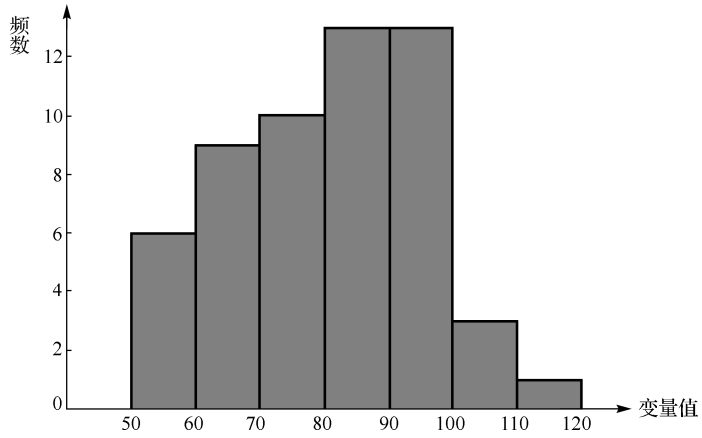


图 5-6　直方图

如果绘制直方图使用的是频数资料，这种直方图是频数直方图，若使用的是频率资料绘制的直方图则称为频率直方图。频率直方图中各个长方形的面积与全部面积之和的比，等于各组的频率，频率直方图与频数直方图的形状完全一致。但是频率直方图中长方形的高度小于 1 或 100%，这对确定长方形高度的刻度比较有利，同时频率相对比较稳定，进行直接的比较不受样本容量大小的影响。

绘制直方图时需要注意的是，直方图是用于展示数值型数据分布特征的工具，直方图中邻近的长方形是相互连接的，不像条形图和直方图在邻近组的长方形之间没有自然的间隔，这是直方图的一般规定。同时与条形图不一样，条形图中的长方形的宽窄没有特别意义，而直方图中长方形的宽度则表示各组的组距。一般，直方图中各长方形的宽度相等，即各组的组距相同，但是若为实现某种特殊目的而编制的频数分布表是不等组距的，各组频数或频率的大小与组距的长短有一定的关系，这时要想绘制直方图以真实地反映数据分布情况，需要将纵坐标改为频数(或频率)密度，即：

$$\text{频数密度} = \frac{\text{组频数}}{\text{该组组距}}$$

二、折线图

折线图又称为多边形图，它是把直方图中各长方形顶端的中点顺次用线段连接起来，得到的表示频数(或频率)分布情况的一种统计图。绘制直方图的准备工作是编制频数分布表，而绘制折线图的基础是直方图。有了直方图之后，只要把直方图各长方形顶端中点标出来，然后用线段连接起来即可，如图 5-7 所示。

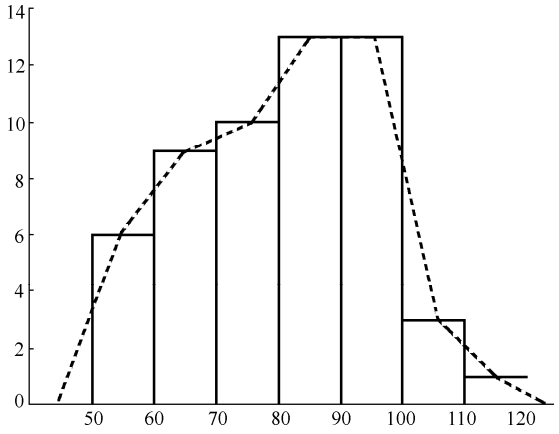


图 5-7　折线图

绘制的折线图在描绘频数分布的轮廓和特征时，看上去比直方图更加简洁明了，尤其是在对截面资料进行多重比较时，用折线图比直方图会更加清晰。如下面两个数据集合的数据分布折线图，如图 5-8 所示。

从图 5-8 可以看到，两个数据集合的分布有明显的差异，其中以虚线为直方图基础的折线图反映出数据主要集中在 60～70 之间，而以实线直方图为基础的折线图反映出数据主要集中在 80～100 之间。进行比较时，显然折线图的对比功能要强于直方图。

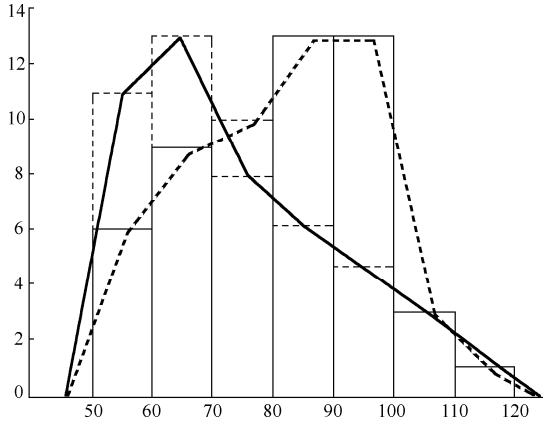


图 5-8　两数据集合折线图比较

三、曲线图

统计实践中，反映数据分布特征的用的更多的是曲线图，曲线图是用一条光滑的曲线近似地描绘频数分布折线图，根据图 5-6 绘制的曲线图如图 5-9 所示。

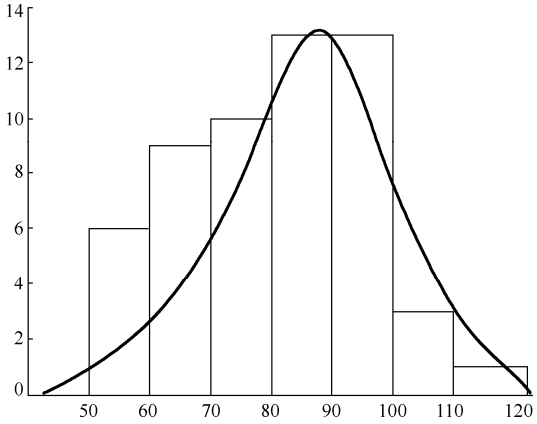


图 5-9　曲线图

四、茎叶图

在统计数据的整理和描述过程中，进行统计分组、编制频数分布表、绘制直方图、折线图和曲线图是使用得比较普遍的传统方法。这些图表在清理统计数据、提炼统计信息方面发挥着重要的作用。但是上面这些图形在反映数据分布特征的同时，也存在着一定的局限，如当原始资料被整理成频数分布表、绘制直方图之后，在表和图中已经看不到原始数据资料的信息，其在反映分布特征的同时缺乏数据资料的还原能力。

根据分组之后的数据资料反映数据的分布特征，我们可以采用直方图、折线图和曲线图，但是如果要用统计图反映未经分组的、且数据较少的原始数据的分布特征，则可以考虑采用茎叶图。

所谓茎叶图是指把每个观察数据划分为两个部分，并分别用茎和叶表示，然后把数据的茎按从小到大的顺序排列，再在每个茎的后面依次列出与之相应的数据的叶的部分，这样便可得到茎叶图。

例如，有下面 50 个数据(如表 5-3 所示)：

表 5-3 数据资料表

59	73	87	65	89	85	77	94	69	97
56	80	68	95	96	50	63	88	91	90
96	92	93	79	74	65	74	89	83	51
74	79	94	67	102	92	100	108	87	86
54	87	86	112	62	76	86	110	86	70

绘制成茎叶图，如图 5-10 所示：

茎	叶												
5	9	6	0	1	4								
6	5	9	8	3	5	7	2						
7	3	7	9	4	4	4	9	0	6				
8	7	9	5	0	8	9	3	7	6	7	6	6	6
9	4	7	5	6	1	0	6	2	3	4	2		
10	0	2	8										
11	0	2											

图 5-10 茎叶图

茎叶图的绘制比较简单，它的主要技巧是设计合适的“茎”，一旦茎确定下来，所有数据就可以固定在相应的各个组中，剩下的工作就是在每个观察值茎的后面填补上数字剩余的部分。一般茎放在左边，叶放在右边，茎与叶之间隔开。

茎叶图的主要作用是可以反映未经分组数据的分布特征，图 5-9 显示出 80～89 的数据最多，而 110 以上的数据最少，表现出两头少、中间多的总体特征，若在图中用线将每个茎后的所有的叶框上，就近似为横着的直方图了(如图 5-11 所示)。当然，茎叶图除了可以反映数据的分布特征外，其还保留了原始数据资料，这是直方图无法做到的。

茎	叶												
5	9	6	0	1	4								
6	5	9	8	3	5	7	2						
7	3	7	9	4	4	4	9	0	6				
8	7	9	5	0	8	9	3	7	6	7	6	6	6
9	4	7	5	6	1	0	6	2	3	4	2		
10	0	2	8										
11	0	2											

图 5-11 茎叶图

利用 Excel 无法直接绘制出茎叶图，我们可以借助于 SPSS 软件。利用 SPSS 绘制茎叶图的做法如下。

首先，选择【Analyze】下拉菜单，并选择【Descriptive Statistics-Explore】选项，进入主对话框。其次，在主对话框中将变量选入【Variables】项，单击【Plots】按钮，在对话框中选择【Stem-and-Leaf】项，单击【Continue】按钮回到主对话框后，单击【OK】按钮即可。

五、箱线图

箱线图是由数据集合中的最大值、最小值、中位数、两个四分位数等五个值绘制而成的。它用于表示一组数据的分布特征，反映数据分布是否对称，是否存在离群点等，同时还可以进行多组数据分布特征的比较，这也是箱线图最大的优点。

箱线图的绘制方法是：首先找出一组数据的最大值、最小值、中位数和两个四分位数；然后用两个四分位数画出箱线图中箱体的两条线；再将最大值和最小值用线段相连接，中位数位于箱体的中间，如图 5-12 所示。

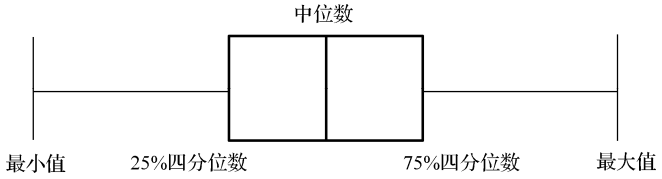


图 5-12　箱线图基本构成

例如：某地按登记注册类型分的城镇单位就业人员平均劳动报酬资料如表 5-4 所示。

表 5-4　按登记注册类型分城镇单位就业人员平均劳动报酬(元)

时间	国有单位	城镇集体单位	股份合作单位	联营单位	有限责任公司	股份有限公司
1	9441	6241	7479	10 608	9750	11 105
2	11 045	6851	8446	11 882	11 024	12 333
3	12 701	7636	9498	12 438	11 994	13 815
4	14 358	8627	10 558	13 556	13 358	15 738
5	16 445	9723	11 710	15 218	15 103	18 136
6	18 978	11 176	13 808	17 476	17 010	20 272
7	21 706	12 866	15 190	19 883	19 366	24 383
8	26 100	15 444	17 613	23 746	22 343	28 587
9	30 287	18 103	21 497	27 576	26 198	34 026

根据上面数据绘制出各类企业就业人员平均劳动报酬，如图 5-13 所示。

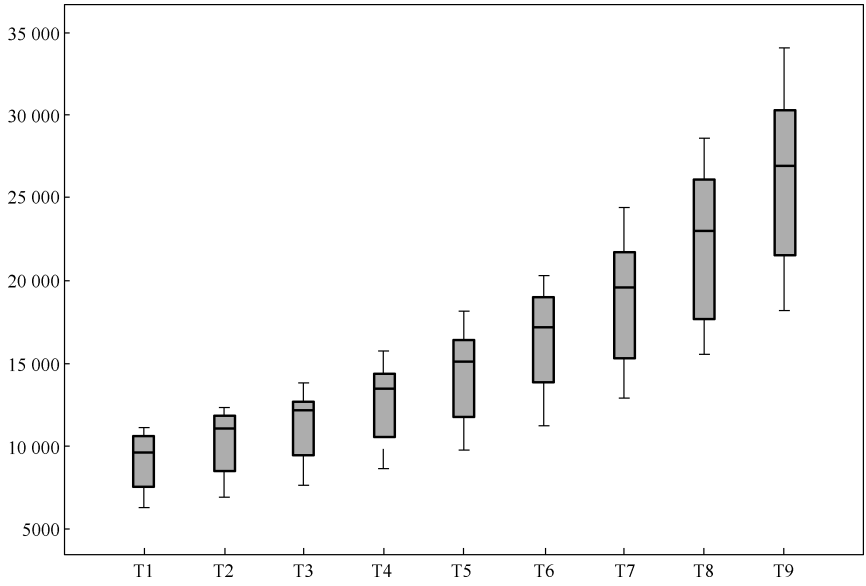


图 5-13　箱线图

图 5-13 显示出了随着时间的进展，各类企业就业人员劳动报酬的中位数水平、最大值、最小值、分位数均逐渐提高，同时平均劳动报酬的最大值与最小值之差也在逐渐拉大。

当然，上面图形显示出的是不同时间上各类企业平均劳动报酬的箱线图，我们也可以用同样的方法绘制关于显示不同类型企业在不同时间纵向上的劳动报酬的箱线图。

六、雷达图

当我们掌握了研究对象的多个指标数据时，我们可以采用柱形图或条形图反映该单位的各指标水平。例如，2016 年某地城镇居民生活消费支出的数据如表 5-5 所示。

表 5-5 2016 年某地居民家庭平均每人全年消费性支出(元)

项 目	总平均	低收入	中等收入	中等偏上	高收入户	最高收入
食品	4259	2846	4181	5044	6087	7874
衣着	1165	599	1136	1433	1857	2643
居住	1145	688	1061	1266	1795	2682
家庭设备用品及服务	691	310	616	838	1210	1926
医疗保健	786	457	749	978	1258	1590
交通通信	1417	533	1079	1634	2633	4986
教育文化娱乐服务	1358	595	1171	1630	2297	3959
杂项商品与服务	418	168	351	495	750	1323

这时，为了直观地反映该地居民家庭平均消费支出，我们可以绘制下面的柱形图(如图 5-14 所示)予以展示：

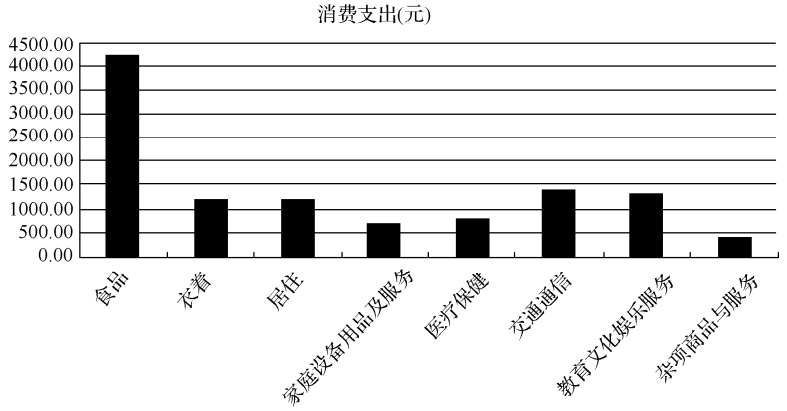


图 5-14 居民家庭消费支出柱形图

但是，如果我们想反映并比较不同收入居民家庭的消费支出时，会涉及五个数据集合(五类收入水平不同的家庭)的八个指标(八类消费支出)，这时可以采用下面的雷达图表示表 5-5 中的数据(如图 5-15)。

雷达图是从一个点出发，用每一条射线代表一个变量，用多个变量的数据点连接成线，即围成一个区域，多个数据集合围成多个区域，即形成雷达图。利用雷达图可以研究多个研究对象在诸多指标之间的相似程度或差异程度。

从图 5-15 中可以看到，该地区居民家庭消费支出中，食品支出额都是最多的，杂项商品与服务支出最少；最高收入居民家庭的各项支出额普遍高于其他收入水平的家庭；除最高收入家庭外，其他收入水平家庭的消费支出结构具有很大的相似性。

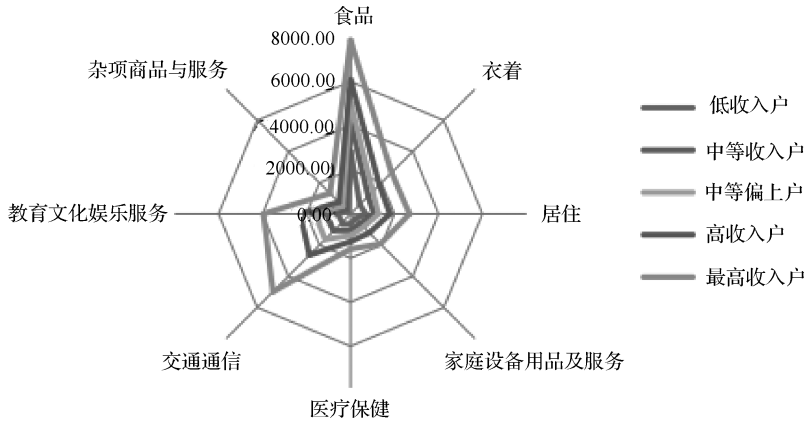


图 5-15　居民家庭消费支出雷达图 单位：元

用 Excel 绘制统计图，可以直接单击【插入】按钮，然后在显示的各类图形中选择我们所需要的图形种类即可。Excel 中的主要图形种类如图 5-16 所示。

图 5-16

第三节　统计图应用中的几个问题

一、合理使用统计图

统计图是描述数据信息的最有效的方式之一，统计图可以把数据的特征信息更加清楚地、直观地显示出来，其最主要的特点是直观的视觉效果。当然，统计图可以给人以深刻的印象，但是如选择了不合适的图或绘制不当，统计图会产生陷阱，对人产生误导。一般绘制统计图时，要注意以下几点。

(1)要画什么样的图，首先要看数据的类型。如若我们要反映的数据是像职业、教育程度之类的定类、定序数据，并且要展示其数据的分布时，就可以采用柱形图；若要展示数据的内部结构，则可以采用饼图；若要表示出一个数值变量如何随着时间改变，就可以用线图，包括折线图和曲线图。当然，如若我们面对的数据是数值型(定距或定比)的数据，而要反映的是其分布，则可以用直方图或茎叶图，在观测值的个数不多时，可以采用茎叶图，而数据资料多时可以采用直方图。

当然，看一个图的时候，要寻找整体形态，以及是否有异于整个形态的异常值。要描述直方图或茎叶图的整体形态，可以看其形状、中心或偏离度。有些分布有简单的形状，比如说是对称或偏斜，但有些分布不太规则，就无法用一个简单的形状来展示。

(2)注意统计图坐标轴的刻度。用同一数据集合可以绘制出不同的统计图，给人以不同的感觉，会得到不同的结论，而其中的奥秘很可能在于图中坐标刻度的确定。如绘制时间数列的折线图或曲线图时，如果将纵轴拉长或缩短，其绘制的统计图的效果会有很大的不同。折线图或曲线图没有所谓正确的刻度，通过对刻度的选择，同样都是正确的图形可以给人很不同的印象，所以要注意统计图中的刻度。

(3)绘制统计图时，要在标识和说明里清楚地表示出图里面画的是什么，单位是什么等

信息，要让数据很醒目。我们需要统计图的美观，但也要注意让看图者的注意力集中在数据本身，而不是标识，也不是背景的图样。我们绘制的统计图是一个呈现数据的图，而不是在从事艺术创作，绘制图形时，应避免一切不必要的修饰，图形所体现的视觉效果应与数据所体现的事物特征一致。

(4)注意不同类型统计图的作用，选用恰当的统计图。如饼图与柱形图相比较，饼图的好处是可以让我们看到数据集合内部的构成特征，其各部分之和为 100%，但是由于角度比长度难比较，因而用饼图中扇形或扇形的角度去比较各部分的大小则不是一个最好的办法。反映不同学历人群所占比重的多少和构成的特征要用饼图，若要比较谁多谁少，则采用柱形图更加合适(如图 5-17 所示)。

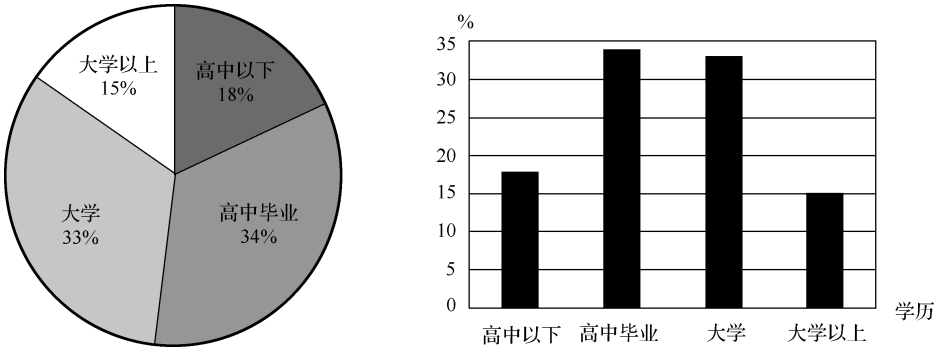


图 5-17　不同学历人群的所占百分比的饼图与柱形图

(5)柱形图是通过比较代表各数量的柱形的高度，来比较各种数据的多少与各指标的大小。但是我们眼睛所看到的，除了高度外还有面积。当所有长条的宽度一样时，面积(高度乘宽度)和高度成正比，所以我们眼睛接收到的信息是正确的印象，因而我们画柱形图时，每个长条都要一样宽，如图 5-17 右图所示。

当然，从艺术美感的角度看，柱形图比较单调，因而有时可能需要用象形图表示(如图 5-18 所示)，象形图其实就是柱形图，只是以具体的图形代替了柱形。但是，要特别留意的是，与柱形图的宽度都一样不同，象形图高度提高的同时，往往其宽度也随之加大，这样，当高度增加一倍时，图形的宽度与高度同时加大，因而给人的视觉就不是大了一倍，而是大了很多，这就容易给人以误导，让人产生错觉。

图 5-18　象形图

二、统计图的绘制实现

绘制统计图，我们可以采用 Excel，也可以采用 SPSS 等多种软件。

1. 采用 Excel 绘制统计图

首先，在 Excel 中单击【插入】菜单，并选择具体的图形，如图 5-19 所示；然后，在各种图形中再进一步选择即可，如图 5-20 所示。

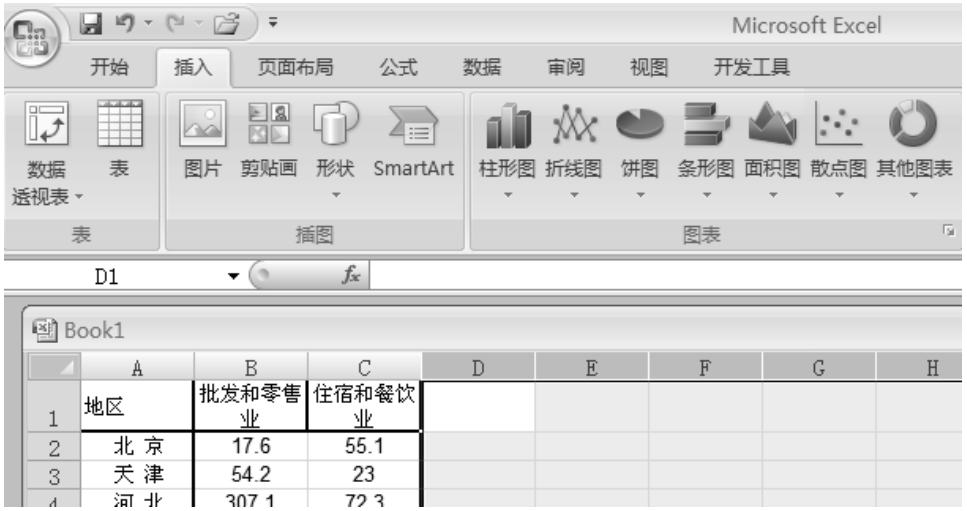


图 5-19　Excel 操作截图 1

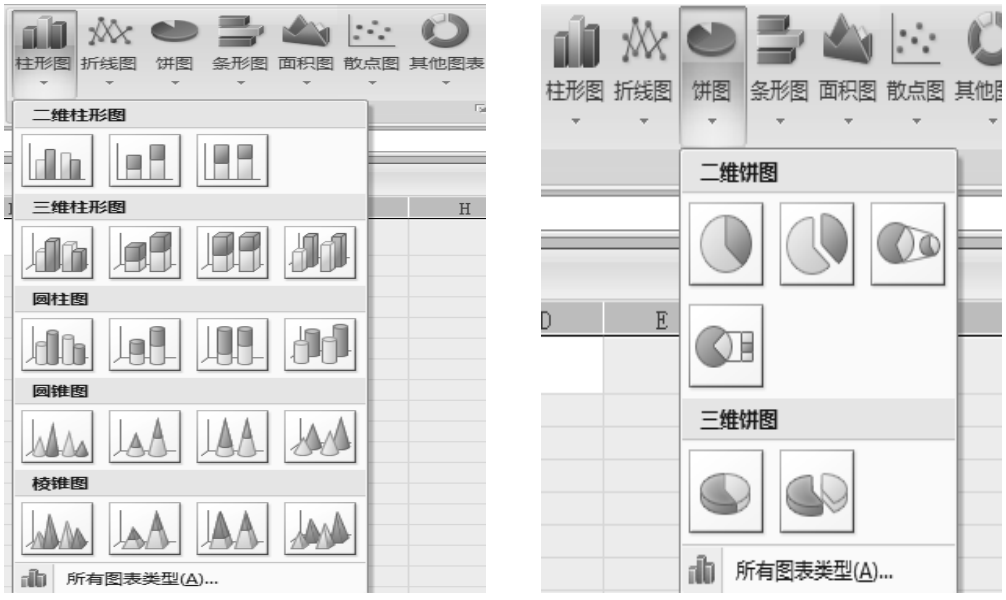


图 5-20　Excel 操作截图 2

2. 采用 SPSS 软件绘制统计图

打开 SPSS 后，单击 Graphs 下拉菜单，在各种选项中选择相应的图形即可。如柱形图

可以选择“Bar”选项；线型图选择“Line”选项；盒形图选择“Boxplot”选项；散点图选择“Scatter/dot”选项；直方图选择“Histogram”选项，等等。

在各图形选项中按软件提示继续输入具体要求即可。

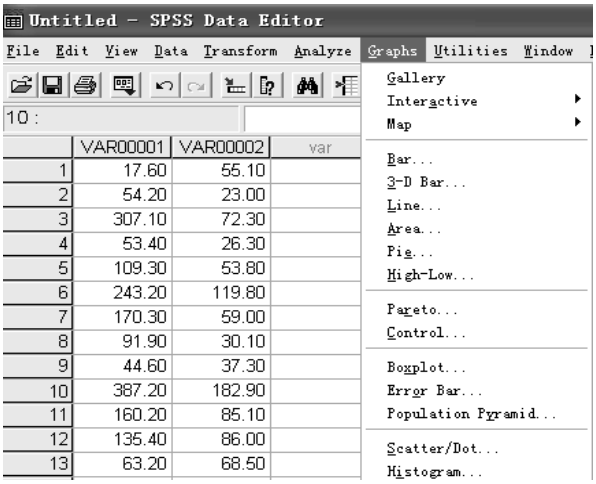


图 5-21 EXCEL 操作截图 3

思考与练习

1．总结各种常用的统计图适用的数据类型。

2．总结各种常用的统计图的作用。

3．要反映数值型数据分布特征可以选用什么样的统计图？

4．已知下面统计资料，利用统计图的方法对数据进行展示，并写出你从图中看到了什么现象。

序号	性别	统计学课成绩	论文成绩	序号	性别	统计学课成绩	论文成绩
1	男	92	优秀	18	男	74	中等
2	男	75	良好	19	男	92	优秀
3	女	62	合格	20	女	90	中等
4	女	90	中等	21	男	76	中等
5	女	62	合格	22	男	62	合格
6	女	84	良好	23	女	63	中等
7	女	92	优秀	24	女	77	中等
8	男	84	中等	25	女	81	优秀
9	男	74	良好	26	女	88	中等
10	男	86	优秀	27	女	92	优秀
11	男	71	合格	28	男	82	中等
12	女	95	良好	29	男	86	良好
13	女	72	中等	30	男	71	中等
14	女	84	中等	31	女	83	优秀
15	女	81	优秀	32	女	83	中等
16	女	75	合格	33	女	71	良好
17	女	76	良好	34	女	88	良好
				35	女	93	良好

第六章　统计指标与多指标综合评价

第一节　统计指标概述

一、统计指标的概念

社会、经济活动的统计描述与分析的重要方法之一就是综合指标法或统计指标(简称指标)法。统计指标是反映同类社会经济现象总体综合数量特征的范畴及其具体数值，是认识事物的基本工具和基本手段，贯穿在统计研究的整体过程中。

统计指标构造的过程包括以下几个阶段。

第一阶段是理论概念在观念上的量化过程，即从一定的理论前提出发，建立观念形态的统计指标和统计指标体系，这一阶段的主要任务是统计指标的设计。

第二阶段是统计指标由观念形态向初级对象形态转化的过程，即统计指标的测量阶段。这一过程表现为前后有机联系的一整套的关于统计调查和统计整理的实际操作。

第三阶段是统计指标由初级对象形态向高级对象形态转化的过程，即统计分析过程。这时的统计指标或指标体系已经成为表现事物统计规律的具体数量或具有某种特定关系的数值体系。

统计指标的构造也可以理解为由观念形态的统计指标向对象形态的统计指标的转化，这包括由观念形态向初级对象形态的转化和初级对象形态向高级对象形态的转化两个阶段。

实际应用中人们对统计指标的理解和使用分为广义和狭义两种。所谓广义的指标的形成包括了统计认识的全过程，而狭义的指标形成只包括前面两个阶段，即观念形态统计指标与统计指标的测量(统计指标向初级对象形态的转化)。本节中所涉及的统计指标主要针对的是狭义的统计指标。

二、统计指标的形成

狭义统计指标的形成分为两个阶段。

(一)观念形态统计指标的形成

统计研究往往是在一定的实质性科学理论指导下进行的。事实上，统计分析与研究的过程很多是从理论概念的量化开始的，理论概念的量化成果就是统计指标及统计指标体系，建立和形成统计指标及统计指标体系是统计研究的第一步，处于这一过程中的统计指标表现为一种观念形态，它是关于某个观念总体数量特征的概念和范畴。

建立或形成统计指标及统计指标体系的过程，就是理论概念的量化过程。统计指标要根据研究目的充分揭示理论概念对象的数量特征，离开了数量化也就无所谓统计指标了。

在量化理论概念形成统计指标的过程中，必须遵循两个基本原则。第一，统计指标及指标体系必须准确地反映理论概念的对象，统计指标对象必须与理论概念的对象相适应，否

则，统计指标将与其理论基础分离，使统计分析研究没有理论的依据。第二，统计指标必须具有由观念形态向对象形态转化的功能，具有实现这种转化的基本条件，否则统计分析与研究将无法继续下去。

要准确地反映理论概念的对象，使统计指标对象与理论概念对象相适应，必须处理好统计指标理论概念的抽象性与具体性之间的关系。理论概念是对统计对象高度概括和抽象的结果，建立和形成统计指标的过程，就是将抽象结果还原为具体的过程。所谓“还原为具体”不是简单地指方向上与抽象过程完全相反，而是指统计学意义上的“还原为具体”。这一过程具有“数量性”和“相对性”两个特点。所谓数量性，是指其是一个量化的过程。在具体化过程中，主要是抓住概念对象的数量化特征。而相对性所指的“具体”是相对于“抽象的理论概念”而言。事实上，作为观念形态的统计指标，其形成过程，既是一个理论概念的具体化过程，又是一个对现实事物数量特征的概括和抽象过程。没有对现实事物的概括和抽象，观念形态的统计指标是无法形成的。

理论概念的抽象是对事物多样性的舍弃，统计指标将其量化、具体化，则是对多样性的一种还原。统计指标具有的特征往往是以统计指标体系的形式存在。例如，“社会产品”作为一个理论概念，唯一表明的是“社会生产活动的成果”，是对这一“成果”多样性舍弃的结果。而要建立社会产品的统计指标，需将其量化、具体化，这时就发现了所谓的“成果”具有多样性。如有总成果，也有净成果，还有最终成果；有物质性成果，也有生产服务性成果；有实物量成果，也有价值量成果，等等。因此，我们需要通过建立社会产品的指标体系来解决这里所遇到的抽象和具体之间的矛盾。

观念形态下的统计指标应具有向对象形态统计指标转化的可能性，这是保证统计分析与研究过程得以继续进行的必要条件。为此，指标设计时必须考虑需要使统计指标具有实际的可测量、可操作性，从而客观上就决定了作为统计指标必须具有的基本要素，即建立和形成统计指标时应当明确规定它的名称和含义，包括内容、计量单位、计算方法等。

应该注意到，不是所有的理论概念都可以直接量化，也不是所有可以量化的概念具有可度量、可操作性。遇到这种情况，我们可以采用一些间接的方法，但是需要考虑统计指标的对象是否和理论概念的对象相适应。

（二）初级对象形态统计指标的形成

形成了观念形态的统计指标后，统计分析与研究过程的下一阶段就是统计指标的测量，也就是取得统计指标的具体数值。所谓测量，是指一个实际的操作过程，其实质在于实现统计指标由观念形态向对象形态的转化，也就是使研究主体观念中的指标概念，通过实际的操作过程变成反映具体对象在具体时间、地点、条件下数量状态的指标数值。在这一阶段形成的对象形态的统计指标，还是一种初级形式。

观念形态统计指标向初级对象形态统计指标的转化过程，也是一个具体化的过程，因此同样需要处理多样性的问题。与形成观念形态统计指标不同的是，在这一阶段，处理多样性问题的原则是坚持统一性。这是对象形态统计指标由初级形式向高级形式转化的重要条件。为此需要有规范的、多样性的各种标准，表现为各种目录，如在社会经济领域内的经济类型目录、行业目录、产品目录、设备目录、材料目录等。

初级对象形态统计指标的形成是在被称为统计调查和统计整理的实际操作过程中实现的，这实际上是一个综合的过程，把总体各单位综合为总体，把表现各单位某一特征的数量

方面的数据综合为总体指标。这一过程具体表现为通过大量观察，记录每一单位的数量特征表现，经过各种分组，汇总出所有单位或各组的单位数及数据总量，以这一总量指标为基础，计算、衍生出其他的派生指标。

在初级对象形态指标的形成过程中，关键的问题是认真地对待和处理指标数值与具体对象真实数据的差异。按照统计误差的分类，这种差异可分为系统性的偏差和随机性误差两类。而在初级对象形态统计指标形成的过程中，要尽可能的消除系统性的偏差，并正确、合理地认识及处理随机误差。

统计分析与研究第二阶段形成的统计指标仅是初级形式的对象形态的统计指标，我们还需要向它的高级形式——高级对象形态统计指标转化。这一转化只有经过统计研究的实证阶段，即统计分析阶段才能完成。当高级对象形态统计指标形成后，也就意味着作为一个具体的统计分析与研究过程基本完成。

三、统计指标的主要类型

统计指标有不同的种类，按指标的表现形式可分为总量指标、相对指标和平均指标三类。

（一）总量指标

总量指标是反映总体数量绝对规模和水平的指标，其数值表现为绝对数，数值的大小和研究对象的总体范围、总体中所包含单位数多少相关。总量指标中有反映一段时期内数量发展过程累计结果的时期总量指标，也有反映某一时刻、某一瞬间、某一时点上数量状态的时点总量指标。其中，时期总量指标是将一段时期中研究对象的数量进行连续登记并累计加总得到的，如企业在 2016 年全年的产值、每个月的产品销售量、企业职工的年度工资总额等。时点总量指标是反映研究对象在某一时点或某一时刻上的数量水平，或将同一时点上各部分数量加总得到的指标，如企业月末的库存量、固定资产价值、地区人口数、银行存款等。

时期指标与时点指标由于所表示的总体数量在性质上不同，因而有不同的特点。时期指标具有可加性，也就是说不同时期的指标数值可以相加，其加总后的数值有明确的实际意义，如可以把 1～3 月每月销售额相加得到第一季度的销售额，其数值大小与时期的长短是有关系的。而时点指标不具有可加性，也就是说在一般情况下不能把不同时点上的数据相加，因为加总后没有实际意义。例如，不能把企业一年 12 个月月末的职工人数相加作为全年的职工人数，也不能把每月月末的库存额相加作为全年的库存，只能反映某一时点或某一时刻上的水平，时点指标数值的大小与时点间隔的大小没有直接的关系。

（二）相对指标

相对指标与总量指标相同，都是对研究对象特征描述的常用的指标之一，它是通过两个有联系的统计指标对比而得到的指标，其具体数值表现为相对数。按指标的特点与作用，相对指标包括结构相对指标、比例相对指标、强度相对指标、动态相对指标、比较相对指标和计划完成程度相对指标等。

1. 结构相对指标

结构相对指标是总体数据集合中某一部分或某几部分数据之和占总体全部数据之和的比重，常称为比重指标。如企业中具有大学本科及以上学历人员所占比重，某地第三产业增加值占地区生产总值的比重等。结构相对指标是描述总体特征的重要指标，可以反映总体内部

构成，是分析研究对象总体分布的基础。其计算公式是：

$$\text{结构相对指标}=\frac{\text{总体某部分数值}}{\text{总体全部数值}}$$

2. 比例相对指标

比例相对指标是研究对象总体中的不同部分数值对比的结果，表明不同部分之间的比例关系，如反映地区产业发展水平，可以用一、二、三产业之间的比例。比例相对指标计算公式是：

$$\text{比例相对指标}=\frac{\text{总体某部分数值}}{\text{总体另一部分数值}}$$

3. 强度相对指标

强度相对指标是将两个性质不同而又有一定联系的指标对比的结果，可以表明事物现象的强度、密度、普遍程度等，如以某种产品产量与人口数对比而得到的平均每人该产品产量，以医院病床数与人口数对比得到的每万人分摊的医院病床数等，其计算公式是：

$$\text{强度相对指标}=\frac{\text{总体某一指标数值}}{\text{另一个有联系的总体指标数值}}$$

在相对指标中，强度相对指标在大多数情况下，都表现为复名数的形式，其单位由分子、分母指标原有的单位组成，如每万人分摊的医院病床数用“医院病床数/每万人”。

4. 动态相对指标

动态相对指标是将某一指标数值在两个不同时间上对比的结果，反映研究对象的发展变化程度，通常也可称为“速度指标”，其对于分析研究对象的发展变化过程十分重要。动态相对指标有发展速度和增长速度两种不同的形式，其计算公式是：

$$\text{发展速度}=\frac{\text{报告期指标数值}}{\text{基期指标数值}}\times 100\%$$

$$\text{增长速度}=\frac{\text{增长量}}{\text{基期指标数值}}\times 100\%$$

5. 比较相对指标

比较相对指标是事物现象某项指标在不同空间或不同场合、不同条件的指标数值对比的结果，表明事物发展的不均衡程度或不同条件下的差异程度，其计算公式是：

$$\text{比较相对指标}=\frac{\text{某一对象某指标数值}}{\text{另一对象的同类指标数值}}$$

例如，生产同种产品的甲乙两企业中，甲企业是乙企业产量的 1.25 倍(或 125%)，甲企业产量比乙企业多 25%。

6. 计划完成程度相对指标

计划完成程度相对指标是一定时期内实际完成的指标数值与计划任务数值对比的结果，一般以百分数形式表示，其计算公式是：

$$\text{计划完成程度相对指标}=\frac{\text{实际完成值}}{\text{计划任务值}}\times 100\%$$

如上所述，相对指标有多种类型，各有自己的意义和作用，计算和应用上也有不同的要求，但从总体上说，在计算和应用相对指标的时候，主要有以下两个基本的要求。

第一，可比性。即要求用以进行对比的指标在指标含义、所包含内容、计算方法等方面要具有一致性或相适应，可以进行对比，且对比结果能够说明问题。

第二，要和总量指标结合应用。如在描述、分析企业生产情况时，在反映总生产规模的同时，可以进一步地通过增长速度反映产量的动态变化，这样可以更全面地认识分析研究对象的现状。

（三）平均指标

在统计指标的各种表现形式中，平均指标具有极其重要的意义，占有极其重要的地位，关于平均指标的计算在本书的相关章节中已有介绍。

应用时需要注意总量指标、相对指标、评价指标的结合应用，以避免对所分析、研究的对象的认识过于片面。

第二节　指标体系与多指标综合评价

一、指标体系

指标可以用于描述、反映研究分析对象某一方面的数量特征，但是任何一个指标都不是全能的，它只能反映研究对象的某一个方面，而不可能描述、反映研究对象的所有方面。实践中，要客观地、全面地反映和描述研究对象，往往需要使用一套指标从多个方面、各个角度进行描述和分析。

指标体系是由一系列相互联系的一整套统计指标构成。统计指标体系对全面深入地认识和研究客观事物现象总体的数量特征与数量关系，揭示总体现象的数量规律有着重要的意义。在很多情况下，它又是推算、估计、评价统计指标的依据。统计指标体系在社会经济领域的数量分析和研究中尤为重要。

通常统计指标体系中各指标之间的联系有以下两种表现形式。

第一种是表现为数学关系式形式。各指标的数量之间表现为加、减、乘、除、乘方、开方等初等数学关系及各种高等数学关系的某种平衡式。例如，商品销售与库存中的平衡关系式：

$$\text{期初库存}+\text{本期购进}=\text{本期销售}+\text{期末库存}$$

又如，生产中劳动投入与产出之间的关系式：

$$\text{总产值}=\text{劳动生产率}\times\text{职工人数}$$

第二种是这些指标之间的关系并不表现为一定的数学关系式，但它们是从不同方面、不同角度去描述研究对象的数量特征，如净资产收益率、总资产报酬率、总资产周转率、流动资产周转率、资产负债率、已获利息倍数、销售增长率、资本积累率等指标构成了对国有企业绩效考核的基本指标体系。而资本保值增值率、主营业务利润率、盈余现金保障倍数、成本费用利润率、存货周转率、应收账款周转率、不良资产比率、现金流动负债比率、速动比

率、三年资本平均增长率、三年销售平均增长率、技术投入比率等构成了国有企业绩效考核的修正指标体系。

二、多指标综合评价方法

所谓多指标综合评价是一种从多个方面利用一套指标体系对研究对象进行评价的统计方法。该评价方法首先是利用可以用于反映研究对象各个方面、具有有代表性的指标进行单方面的评价，然后在此基础上进行多指标综合，从而达到对研究对象进行全面、客观的评价。

多指标综合评价需要解决下面几个关键问题。

(一)构建用于评价的指标体系

评价指标体系就是可以综合反映评价目的的、具有代表性的和客观评价研究对象的指标体系。构建评价指标体系，首先需要进行理论研究，其中包括统计指标理论及统计指标体系的理论研究，以便为确定所需的评价指标提供一定的理论依据。构建的评价统计指标体系是否合理，直接关系到对评价对象评价结果的科学性和客观性。指标体系的建立，应进行必要的定性研究，对所研究的问题进行深入的分析，尽量选择那些能够满足评价目的、具有一定综合意义、具有代表性、具有可操作性的指标；指标体系的建立还需要尽可能地运用一定的统计方法进行筛选，避免信息的重叠，以提高指标的客观性。

中央企业绩效考核评价的指标体系如下：

评价内容	基本指标	修正指标	评议指标
财务效益状况	净资产收益率 总资产报酬率	资本保值增值率 主营业务利润率 盈余现金保障倍数 成本费用利润率	经营者基本素质 产品市场占有能力 (服务满意度) 基础管理水平 发展创新能力 经营发展战略 在岗员工素质 技术装备更新水平 (服务硬环境) 综合社会贡献
资产营运状况	总资产周转率 流动资产周转率	存货周转率 应收账款周转率 不良资产比率	
偿债能力状况	资产负债率 已获利息倍数	现金流动负债比率 速动比率	
发展能力状况	销售(营业)增长率 资本积累率	三年资本平均增长率 三年销售平均增长率 技术投入比率	

(二)确定评价的标准，进行单指标评价

对研究对象进行评价，需要存在一个评价的标准。评价的结果可以有多种，如百分制的、五分制的等，而每个评价等级需要确定该等级的标准。不仅仅是综合评价等级的标准，在对单项指标进行评价时也需要有单指标评价等级的标准，有了标准才可以对研究对象进行评价。在单指标评价时，评价的标准需要结合评价的目的来确定，应用时常用的评价的标准有以下几种。

(1)动态标准，即时间标准。可以某固定时期水平作为评价的标准，可以用历史上最好水平为评价标准，也可以上一期作为评价标准，等等。

(2)静态标准，即横向标准。比如，以行业平均水平作为标准、行业最高水平(或最低水平)作为标准等。

(3)理论标准，即使整个系统达到最佳或达到某种状态下的水平作为评价等级的标准。

(4)主观标准，即以评价者的主观感受作为评价标准。

(5)其他。

中央企业绩效考核评价中，绩效定性评价标准分为优(A)、良(B)、中(C)、低(D)、差(E)五个档次。其中，企业财务绩效定量评价标准值的选用，是根据企业的主营业务领域对照企业综合绩效评价行业的基本分类，自下而上逐层遴选被评价企业适用的行业标准值。多业兼营的集团型企业财务绩效指标评价标准值的选用区分了主业突出和不突出两种情况。(1)存在多个主业板块但某个主业特别突出的集团型企业，采用的是该主业所在行业的标准值。(2)存在多个主业板块但没有突出主业的集团型企业，是根据其下属企业所属行业，分别选取相关行业标准值进行评价，然后按照各下属企业资产总额占被评价企业集团汇总资产总额的比重，加权形成集团评价得分；也可以根据集团的经营领域，选择有关行业标准值，以各领域的资产总额比例为权重进行加权平均，计算出用于集团评价的标准值。

各指标的各评价等级的标准确定后，就可以进行单指标的评价。单指标的评价的目的是也为了进行多指标的综合，因为不同的指标具有不同的量纲，要想综合成为一个指标，则需要对各指标进行无量化的处理，即进行单指标的评价。单指标评价方法主要有直线法、折线法和曲线法。

所谓直线法是指指标数值与指标的评价结果表现出一种直线的关系，而折线法指标的评价结果表现出一种折线的关系，即两者之间的关系不是固定比例的关系。

(三)确定各指标的权重

由于各指标在综合评价时的作用、重要性与意义不同，在对各指标的单指标评价结果进行综合前，需要确定各指标的权重。权重的确定方法有多种，大体上可分为两类：一类是主观构权法；另一类是客观构权法。

主观构权法是指研究者根据其主观判断来确定各指标的重要性，即权数的一种方法，通常采用的是专家评判法。客观构权法是相对主观构权法而言的，是直接根据指标的原始信息，通过统计方法处理后获得权数的一种方法。两种方法各有利弊，主观构权法往往没有统一的客观标准，客观构权法可在一定程度上弥补这一不足，实际应用中最好是将两种方法结合使用。

(四)确定各单项指标综合的合成方法，进行单指标评价结果的综合

将各单指标评价结果进行综合时，常采用的方法有加法合成法、乘法合成法等。

(1)所谓加法合成法也就是加权算术平均法，其综合合成的公式是：

$$I=\frac{\sum_{i=1}^{n} z_i w_i}{\sum_{i=1}^{n} w_i}$$

式中，z_i 表示各单指标评价结果；w_i 是各指标的权数；n 是指标的个数。

(2)所谓乘法合成法也就是几何平均法，其综合合成的公式是：

$$I=\sqrt[\sum w_i]{z_1^{w_1}\cdot z_2^{w_2}\cdot\cdots\cdot z_n^{w_n}}$$

实际应用时，可以结合具体的问题将加法合成和乘法合成两种方法结合在一起，即加乘混合法。

思考与练习

1．说明统计指标的形成过程。

2．确定某一研究对象，并结合研究对象举例说明几种总量指标、相对指标、平均指标的作用、计算和应用中应注意的问题。

3．说明指标体系的作用。

4．说明统计综合评价的方法和应用中应注意的问题。

第二部分　单一指标的统计推断

统计推断是在不掌握总体全部数据信息的情况下，利用由样本数据计算的统计量去推断总体的数量特征。统计推断从内容看包括对总体单一指标的推断、统计分布的推断，也包括对总体各变量之间关系的推断等。在推断方法上包括参数估计方法与非参数估计方法等等。下面三章围绕总体的单一指标推断，介绍在一定分布或条件下的参数估计方法，包括抽样分布、参数估计和假设检验。

第七章　概率抽样方法与抽样分布

第一节　随机变量的概率分布

一、随机变量

在投掷一枚均匀骰子的随机试验中，每次投掷的结果可能会出现不同的点数，这个“出现的点数”就是一个变量，它的全部可能取值为 1、2、3、4、5、6。在一次试验中到底出现哪一个点数虽然是随机的，其具体结果不能准确地预料，但是每个取值出现的概率却是确定的，皆为六分之一。这样我们就可以用 “出现的点数”这一变量的所有可能取值以及这些值出现的概率来描述这个随机现象。

随机试验的结果，可以用数据描述，我们称其为随机事件。而用于描述试验结果的变量称为随机变量。如上例中“出现的点数”就是一个随机变量，它的所有可能取值，即数据为 1、2、3、4、5、6。再例如测量一群人的身高，这群人的“身高”就可看作是一个随机变量。如果要观察这一群人的性别，“性别”就是随机变量，其具体数据就表现为男性和女性。观察产品质量是否合格，“产品质量”就是变量，其具体数据表现为“合格”与“不合格”，等等。对于随机变量，我们也可以用 X、Y、Z 等字母表示。这样我们就可以把对随机事件及其概率的研究转化为对随机变量取值及其概率的研究，以便于分析研究对象的数量规律。

按照随机变量的特性，通常我们可以把随机变量分为离散型随机变量和连续型随机变量两类。若随机变量 X 的所有可能取值可以一一列举，即所有可能取值为有限个或无限可列个，则称 X 为离散型随机变量。而如果随机变量 X 的所有可能取值不能逐个列举出来，而其所有可能值为某一区间，例如一批电子元件的使用寿命就是在某一区间范围内，这时称 X 为连续型随机变量。

二、离散型随机变量的概率分布

(一) 离散型随机变量概率分布的表示方法

随机变量所有可能取值及其相应的概率，称为随机变量的概率分布。如对于离散型随机变量 X，设其所有可能取值为 $x_1, x_2, \cdots, x_k, \cdots$，取这些值的概率依次为 $p_1, p_2, \cdots, p_k, \cdots$。作为概率分布其必须满足两个基本条件：第一，变量每个可能取值的概率均为非负；第二，所有可能取值的概率之和必须等于 1。

一般离散型随机变量概率分布常用的表示方法主要有三种。

(1) 公式法：$P(X = x_k) = p_k \quad (k = 1, 2, \cdots)$。

(2) 列表法(如表 7-1 所示)。

表 7-1　离散型随机变量概率分布

X	$x_1, x_2, \cdots, x_k, \cdots$
P	$p_1, p_2, \cdots, p_k, \cdots$

(3)图示法：以横轴代表 X 取值，纵轴代表概率 p，将随机变量的取值和相应的概率形成的坐标点绘制在直角坐标系上，并将这些点与相应的横坐标上的 X 连接起来，即是离散型随机变量的概率图(如图 7-1 所示)，通常称 $P(X)$ 为随机变量 X 的概率函数。

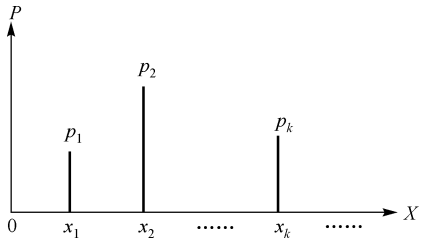


图 7-1　离散变量的概率图

从离散型随机变量的概率分布中，我们不仅可以知道它取各个可能值的概率，而且还可以求出它在某范围内取值的概率，所以离散型变量的概率分布较好地描述了相应的随机试验。

(二)离散型随机变量的期望值与方差

随机变量的概率分布包含了随机变量概率性质的一切信息，但在实践中我们往往需要对随机变量的分布特征进行概括性的度量，如变量的中心位置及离散程度等。为了获得概率分布的中心位置及变量变异程度或离散程度，我们需要计算随机变量的期望值和方差。

1. 离散型随机变量的期望值

例 7-1：某台设备一段时间内发生故障的次数 X 及其相应的概率如表 7-2 所示。

表 7-2　故障次数的概率分布表

故障次数 X	0	1	2	3
概率 P	0.65	0.21	0.12	0.02

那么，该设备这段时间内平均发生故障的次数应为：

$$0\times0.65+1\times0.21+2\times0.12+3\times0.02=0.51\text{(次)}$$

这就是离散型随机变量的数学期望。

随机变量的期望值记为 $E(X)$，其计算公式是：

$$E(X)=\sum_i x_i p(X=x_i)=\sum_i x_i p_i$$

一个随机变量的数学期望是对该随机变量概率分布中心位置的度量，它反映了随机变量的平均值。

描述统计中的平均数与随机变量的数学期望具有相似的性质和作用，即都是反映集中趋势的指标。但描述统计中的平均数是对某一组观察到的具体数据而言，而随机变量的数学期望是对某一总体抽象分布而言的，它代表了该类数据的总体水平，反映的并不是已观察到的数据，而是假设潜在可能发生的数据。

2. 离散型随机变量的方差及标准差

随机变量的方差是用来反映随机变量取值的离散程度的。随机变量 X 的方差定义为 X 每一个取值与其期望值离差平方之期望值，记为 $D(X)$。

离散型随机变量方差的计算公式是：

$$D(X)=E[X-E(X)]^2=\sum_i[x_i-E(X)]^2\cdot p_i=E(X^2)-(EX)^2$$

由方差的定义可知，若 X 的取值比较集中，则其方差较小；若 X 的取值比较分散，则其方差较大。

将 X 方差 $D(X)$ 开平方，称为随机变量 X 的标准差。

例 7-2：某企业在其销售的产品包装内设置了抽奖的项目，其中一等奖 8%，奖金 50 元；二等奖 10%，奖金 30 元；三等奖 15%，奖金 10 元。求任意购买一件商品获奖金额的数学期望和方差。

解：X 所有的取值为 50、30、10、0，取这些值的概率 P 分别为 0.08、0.10、0.15 和 0.67。

计算的数学期望：

$$E(X)=\sum_i x_i p_i=50\times0.08+30\times0.1+10\times0.15+0\times0.67=8.5$$

即每购一件商品，所中奖的平均奖金额为 8.5 元。

$$E(X^2)=\sum_i x_i^2 p_i=50^2\times0.08+30^2\times0.1+10^2\times0.15+0^2\times0.67=305$$

则 X 的方差为

$$D(X)=E(X^2)-(EX)^2=305-8.5^2=232.75$$

（三）常用的离散型随机变量的概率分布

1. 两点分布

两点分布也称为 0～1 分布。设随机变量 X 只可能取 0 与 1 两个值，它们的概率分布是：$P(X=k)=p^k(1-p)^{1-k},k=0,1$，则称 X 服从参数为 P 的两点分布。其数学期望值为 p，方差为 $p(1-p)$。

2. 二项分布

实际问题中，有许多试验的所有可能结果只有两种，要么是“是”，要么是“否”，这时如果试验若干次，则结果为“是”的次数就是一个随机变量。与这类实验相似的具有下面特征的试验称为贝努里试验：(1)试验包含了 n 次相同的试验；(2)每一次试验结果只有两个可能的结果，即“是”或“否”、“成功”或“失败”等；(3)每次试验出现“是”的概率均为 p，出现“否”的概率为 $1-p$；(4)每次试验是相互独立的；(5)试验结果为“是”或“否”是可以计数的，即试验结果对应一个离散型随机变量。

以 X 表示 n 次重复独立的贝努里试验中结果“是”出现的次数，则该次数服从二项分布，$X=k$ 的概率为：

$$P(X=k)=C_n^k p^k(1-p)^{n-k}\qquad(k=0,1,2,\cdots,n)$$

从上式中可以看出，当 n=1 时，二项分布就转化为两点分布，所以两点分布是二项分布的特殊情况。二项分布是两点分布的推广，在实践中有广泛的应用。

可以证明，服从二项分布的随机变量其期望值和方差分别为：

$$E(X)=np\ ,\quad D(X)=np(1-p)$$

标准差为：$\sqrt{D(X)}=\sqrt{np(1-p)}$。

例 7-3：若某企业有 80%的职工的工资超过了平均工资，现从所有职工中采用重复抽样的方法抽取 20 人，则 20 人中恰好有 16 人的工资超过平均工资的概率是多大？

解：我们把抽取一个职工进行调查作为一次试验，由于是采用重复抽样的方法，抽取 20 个职工可以看成是进行了 20 次独立的试验，抽取每一个职工其工资超过平均工资的概率为 0.8，则抽取 20 人恰好有 16 人工资超过平均工资的概率为：

$$P(X=16)=C_{20}^{16}0.8^{16}(1-0.8)^{20-16}=0.218$$

而 20 人中工资超过平均工资的人数平均为：

$$E(X)=np=20\times0.8=16$$

方差为：

$$D(X)=np(1-p)=20\times0.8\times0.2=3.2$$

3. 泊松分布

泊松分布是很常见的一种分布，许多随机现象都服从泊松分布。例如，一匹布上疵点的个数；书中一页纸上印刷错误的个数；某段时间内仪器出现故障的次数，等等。

泊松分布常用来描述在一指定时间范围内或在指定的面积、体积之内某一事件出现的次数的分布。服从泊松分布的随机现象主要集中在社会生活和物理学领域中。在社会生活中，又尤其适用于各种对服务的需求现象或排队现象。

设随机变量 X 可取无穷多个值 0, 1, 2, …，其概率分布为：

$$P(X=k)=\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}\qquad(k=0,1,2,\cdots,\lambda>0)$$

则称 X 服从参数为 λ 的泊松分布。

其中，λ 的含义为给定的时间间隔内事件的平均数。可以证明：泊松分布的数学期望与方差均为 λ，即 $E(X)=\lambda,\ D(X)=\lambda$。

例 7-4：假定某企业职工在周三请事假的人数 X 近似服从泊松分布，且周三请事假的平均人数为 2.5 人。要求：(1)计算周三请事假的人数 X 的期望值和标准差；(2)在给定的某周三正好请事假的人数为 5 人的概率。

解：(1) X 的期望值= $E(X)=\lambda=2.5$

$$X\text{的标准差}=\sqrt{D(X)}=\sqrt{\lambda}=\sqrt{2.5}\approx1.58$$

(2) $P(X=5)=\dfrac{2.5^5\mathrm{e}^{-2.5}}{5!}=0.066801$。

三、连续型随机变量的概率分布

(一)概率密度函数与分布函数

1. 概率密度函数

由于连续型随机变量可以取某一区间或整个实数轴上的任意一点，所以我们不能像对离

散型随机变量那样列出每一个变量值及其相应的概率，只能采用概率密度函数。当连续型随机变量分布曲线图可以用数学函数形式 $f(x)$ 表示时，称 $f(x)$ 为概率密度函数。

概率密度函数应满足下面两个条件：

(1) $f(x)\geqslant 0$；

(2) $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$。

需要指出的是 $f(x)$ 并不是一个概率，而是一个函数，$f(x)\neq P(X=x)$。连续型随机变量取任何一个个别值的概率为零，用分布数列或通项公式不仅不可能描述一个连续型随机变量，而且也是无意义的，这是与离散型随机变量不同的一个重要特征。另外，由于连续型随机变量取任一值的概率等于零，所以连续型随机变量在任一区间上取值的概率与是否包含区间端点无关，即：

$$P(a<X<b)=P(a<X\leqslant b)=P(a\leqslant X<b)=P(a\leqslant X\leqslant b)=\int_a^b f(x)\mathrm{d}x。$$

连续型随机变量 X 在 a 与 b 之间的概率可以表示为概率密度函数 $f(x)$ 在 $[a,b]$ 区间上的曲线下方的面积，如图 7-2 所示。

当连续型随机变量的概率密度函数确定后，可以通过定积分求得连续型随机变量在任一区间上取值的概率，所以连续型随机变量的概率密度描述了相应的随机试验。

2. 分布函数

连续型随机变量也可以用分布函数 $F(x)$ 来表示，分布函数定义为：

$$F(x)=P(X\leqslant x)=\int_{-\infty}^{x} f(t)\mathrm{d}t \qquad (-\infty<x<+\infty)$$

它也是建立在概率密度 $f(x)$ 的基础上的。因此 $P(a<X<b)$ 也可以写成：

$$\int_a^b f(x)\mathrm{d}x=F(b)-F(a)$$

分布函数与概率密度函数的关系是：分布函数对 x 的导数为概率密度，即：

$$F'(x)=f(x)$$

分布函数的概念看起来很抽象，实际上它具有明确的意义。它是一种概率，对任意给定的一个 x，$\{X\leqslant x\}$ 是一个随机事件，而 $F(x)$ 就是这一事件发生的概率。

分布函数具有以下基本性质：

(1) $0\leqslant F(x)\leqslant 1$；

(2) $F(x)$ 是一个单调非减的函数(如图 7-3 所示)。

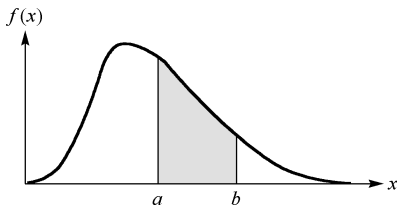


图 7-2　连续型随机变量分布图

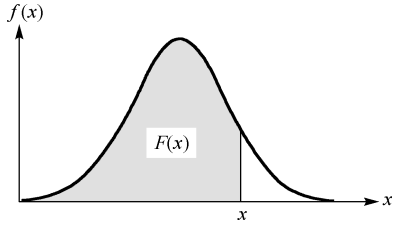


图 7-3　分布函数图

（二）连续型随机变量的期望值与方差

已知连续型随机变量 X 概率密度为 $f(x)$，若 $\int_{-\infty}^{+\infty} xf(x)\mathrm{d}x$ 绝对收敛，则称其为 X 的数学期望，记作 $E(X)=\int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x$。

方差 $D(X)$ 为：

$$D(X)=\int_{-\infty}^{+\infty}[X-E(X)]^2 f(x)\,\mathrm{d}x=E(X^2)-[E(X)]^2$$

（三）常见的连续型随机变量的概率分布

1. 均匀分布

如果随机变量 X 的概率密度函数为：

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & a\leqslant b\leqslant a\\ 0, & \text{其他}\end{cases}$$

则称 X 服从区间[a, b]上的均匀分布（如图 7-4 所示）。

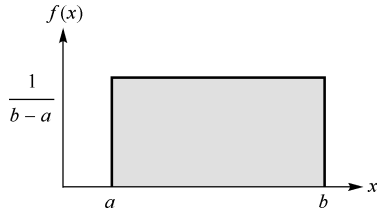


图 7-4　均匀分布图

显然，服从均匀分布的随机变量在其取值范围[a, b]内的概率密度函数是一个常量，也就是说，服从均匀分布的随机变量在其区间内取任何一个相等区间段数值的概率都相同，其在任何小区间上取值的概率的大小只与该小区间的长度成正比，而与该小区间的具体位置无关。

在区间[a, b]上均匀分布的随机变量 X 的数学期望和方差为：

$$E(X)=\frac{a+b}{2},\qquad D(X)=\frac{(b-a)^2}{12}$$

2. 正态分布

正态分布是所有概率分布中最重要的一种分布。实践中，我们观察到的现象有很多都是近似服从正态分布，如产品的长度、宽度、高度；人体的身高、体重；测量的误差，等等，都近似服从正态分布。事实上，如果影响某一随机变量的因素很多，而每一个因素都不起决定性作用，且这些影响因素可以叠加，那么，这一个随机变量就往往被认为是服从正态分布。从理论上讲，正态分布可以导出一些其他的分布，而有些分布在一定条件下又可用正态分布来近似表示，因此正态分布在理论研究与实际应用中均有重要地位。

(1)一般正态分布。

如果随机变量 X 的概率密度是：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (-\infty<x<+\infty)$$

则称 X 服从一般的正态分布，简称正态分布，记作 $X\sim N(\mu,\sigma^2)$。其中：μ 为随机变量 X 的期望值，σ 为 X 的标准差，它们是正态分布的两个参数。

正态分布的概率密度 $f(x)$ (如图 7-5 所示)具有以下性质。

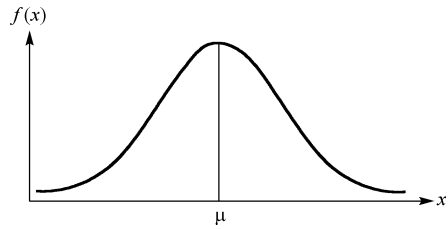


图 7-5 正态分布图

第一，$f(x)\geqslant 0$，即整个概率密度曲线都在 x 轴的上方，且呈钟形(如图 7-5 所示)。

第二，曲线 $y=f(x)$ 关于直线 $x=\mu$ 对称，并在 $x=\mu$ 时达到极大值，$f(x)=\frac{1}{\sigma\sqrt{2\pi}}$。

第三，μ 决定了图形的中心位置，曲线的陡缓程度由 σ 决定。σ 越大，X 方差越大，曲线越平缓；σ 越小，X 方差越小，曲线越陡峭(如图 7-6 所示)。

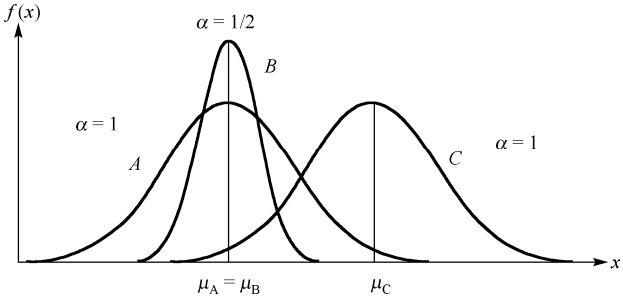


图 7-6 正态分布图

第四，当 x 趋于无穷大时，曲线以 x 轴为其水平渐进线。

(2)标准正态分布。

如果正态分布的期望值为 0，方差为 1 时，则 $f(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ 称为标准正态分布，记为：$X\sim N(0,1)$。

对于标准正态分布，通常用 $\phi(x)$ 表示概率密度，用 $\Phi(x)$ 表示分布函数，即 $\phi(x)=\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$，$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t$。

标准正态分布的概率密度函数的图形如图 7-7 所示。

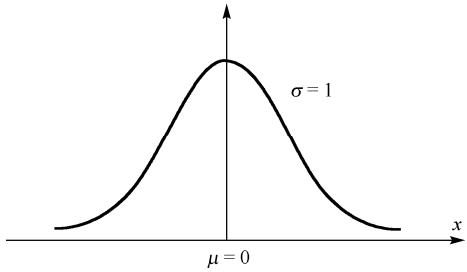


图 7-7　标准正态分布图

标准正态分布的重要意义在于：任何一个一般的正态分布都可以通过线性变换(即标准化变换)转换为标准正态分布。

设 $X \sim N(\mu,\sigma^2)$，可以证明：$Z=\dfrac{X-\mu}{\sigma} \sim N(0,1)$，这就是将一般正态分布转化为标准正态分布的方法。

可以证明：$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$ 及 $\int_{-\infty}^{+\infty} \phi(x)\mathrm{d}x=1$，即正态曲线下的面积为 1。

如果要直接计算一般正态分布曲线下的面积，需要采用积分的方法。除此以外，我们可以将服从一般正态分布的变量 X 转化为服从标准正态分布的变量 Z，然后利用标准正态分布表得到相应的结果(如图 7-8 所示)。

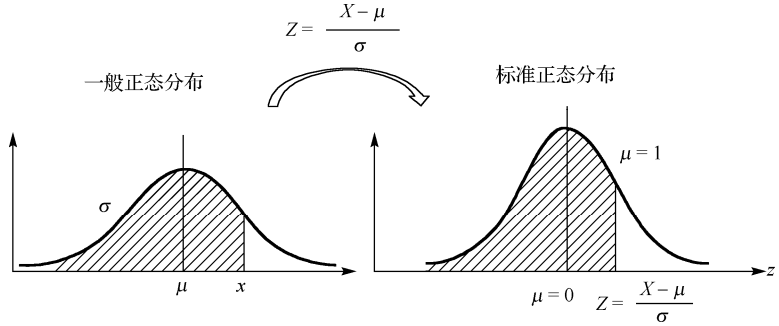


图 7-8　一般正态分布转标准正态分布

对于标准正态分布，由 $\Phi(x)$ 的对称性可知：$\Phi(-x)=1-\Phi(x)$。

若 Z 服从标准正态分布，由于正态曲线下的面积为 1，所以 $\Phi(0)=\dfrac{1}{2}$，查正态分布表可知：

$$P(Z<1.5)=P(-\infty<Z<1.5)=\Phi(1.5)=0.9332$$

则

$$P(Z>1.5)=1-P(Z<1.5)=1-0.9332=0.0668$$

$$P(|Z|<1.5)=P(-1.5<Z<1.5)=\Phi(1.5)-\Phi(-1.5)$$
$$=\Phi(1.5)-[1-\Phi(1.5)]=2\Phi(1.5)-1=2\times 0.9332-1=0.8664$$

对于一般的正态分布，如果要计算其某一区间内的概率，可首先将其转换为标准正态分布，然后按标准正态分布的方式计算即可得到结果。

例 7-5：设 $X \sim N(5,3^2)$，求以下概率：

(1) $P(X \leqslant 10)$；(2) $P(2 < X < 10)$。

解：(1) $P(X \leqslant 10) = P\left(\dfrac{X-5}{3} \leqslant \dfrac{10-5}{3}\right) = P\left(\dfrac{X-5}{3} \leqslant 1.67\right)$

$$= P(Z \leqslant 1.67) = \Phi(1.67) = 0.9525$$

(2) $P(2 < X < 10) = P\left(\dfrac{10-5}{3}\right) - P\left(\dfrac{2-5}{3}\right) = \Phi(1.67) - \Phi(-1)$

$$= \Phi(1.67) - [1 - \Phi(1)] = 0.9525 - [1 - (0.8413)] = 0.7938$$

例 7-6：设某种零件的长度服从正态分布，其平均长度为 10 毫米，标准差为 0.2 毫米，试问：

(1) 从该批零件中随机抽取一件，其长度不到 9.4 毫米的概率是多少？

(2) 为了保证产品质量，要求以 95%的概率保证零件的长度在 9.5～10.5 毫米之间，这一要求能否得到保证？

解：已知 $X \sim N(10,\ 0.2^2)$

(1) $P(X < 9.4) = P\left(\dfrac{X-10}{0.2} < \dfrac{9.4-10}{0.2}\right) = P(Z < -3) = 1 - \Phi(3) = 0.0013$；

(2) $P(9.5 < X < 10.5) = P\left(\dfrac{10.5-10}{0.2}\right) - P\left(\dfrac{9.5-10}{0.2}\right) = \Phi(2.5) - \Phi(-2.5) = 0.9876$。

即可以用 98.76%的概率保证该批零件的长度在 9.5～10.5 毫米之间。

若 X 服从二项分布，可以证明当 n 很大且 $0 < p < 1$ 是一个定值时，$\dfrac{X - np}{\sqrt{np(1-p)}}$ 近似服从 $N(0,\ 1)$，进一步可以得到 $P(a \leqslant X \leqslant b) \approx \Phi\left(\dfrac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\dfrac{a-np}{\sqrt{np(1-p)}}\right)$。这一正态近似很重要，它提供了计算二项分布概率的一种实用、简便的近似方法。

例 7-7：100 台车床彼此独立地工作着，每台车床的实际工作时间占全部时间的 80%，求：

(1) 任一时刻，有 70～86 台车床在工作的概率；

(2) 任一时刻，有 80 台以上车床在工作的概率。

解：将任一时刻观察每台车床是否工作看成是一次试验，100 台车床可看做 100 次试验，每次试验成功(车床工作)的概率 $P = \dfrac{80}{100} = 0.8$。

设 X 表示 100 台车床中工作着的车床台数，其期望值 $E(X) = np = 100 \times 0.8 = 80$，标准差 $\sqrt{D(X)} = \sqrt{np(1-p)} = \sqrt{100 \times 0.8(1-0.8)} = 4$。

于是：(1) $P(70 \leqslant X \leqslant 86) = \Phi\left(\dfrac{86-80}{4}\right) - \Phi\left(\dfrac{70-80}{4}\right) = \Phi(1.5) - \Phi(-2.5)$

$$= \Phi(1.5) + \Phi(2.5) - 1 = 0.9270$$

(2) $P(X > 80) = P(80 < X \leqslant 100) = \Phi\left(\dfrac{100-80}{4}\right) - \Phi\left(\dfrac{80-80}{4}\right) = \Phi(5) - \Phi(0)$

$$= 1 - 0.5 = 0.5。$$

第二节　概率抽样方法

一、基本概念

(一)总体与样本

所谓总体就是研究对象中的所有个体或数据集合，它可以是由具有某种共同性质的许多个体如企业、居民户、职工等组成，也可以是由某一现象的所有个别观察值所组成，这样总体就可以是一个具有确切分布的随机变量 X。

样本是指按照一定的抽样原则从总体中抽取的一部分单位的集合，抽取样本的目的是通过对样本的观察和研究，达到对总体的认识。

总体根据其所包含的总体单位的数目可以分为有限总体和无限总体。所谓有限总体是指总体具有明确确定的范围，而且单位的数目是有限可数的，如所要分析、研究的企业的所有职工，一批待检验的灯泡等。而无限总体是指总体中包含的总体单位数是无限多的，例空气中的颗粒、可以无限延续下去的科学试验所产生的试验结果等。把总体区分为有限总体和无限总体主要是为了判别在抽样过程中逐个抽取个体时，每次抽取是否独立。在无限总体中，抽取的每一个总体单位之间可以看作是独立的，而在有限总体抽取总体单位时，每次抽取不一定是独立的，它可能会受到以前抽取结果的影响。

(二)概率抽样

前面章节中已经提到过概率抽样的概念，为进一步阐述抽样分布的理论，在此再次提到概率抽样。所谓概率抽样是指按照随机原则在总体中抽取样本的方法，这样抽取的样本称作随机样本。在这种抽样方法的基础上，我们可以根据抽样分布的理论与方法，用样本去推断总体，并对所得结果给出一定的置信度。

二、简单随机抽样

简单随机抽样是指对总体不做任何处理，完全按随机的原则，直接从总体全部单位中抽选出样本单位加以观察。简单随机抽样在抽取样本时，完全不受主观意志的支配，从理论上来说是最符合抽样调查的随机原则的抽样方式，是概率抽样的最基本的形式。简单随机抽样的具体操作方法有：直接抽选法、抽签法、随机数字表法等。

1. 直接抽选法

这种方法就是直接从调查对象中随机抽选。例如，从水池中直接抽选一定数量的水样化验；随机地确定商品仓库中不同的地点并取出若干同类商品作为样本进行质量检验等。这种方法一般适用于小型总体，对于大型总体很难实施此法。

2. 抽签法

它是给总体的每个单位进行编码，然后按照随机的原则从中抽选，直到抽够规定的数量为止。在理论上此方法简单，但在实践中，对于总体单位数目有很多的情况下，编码工作量过大，另外也很难将各单位搅动均匀。因此，这种方法也有其局限性。

3. 随机数字表法

它是将 0, 1, 2, …, 9 十个数字，按随机原则编排成数字表，利用随机数字表抽选样本单位。这种方法首先要对总体各单位进行编号，然后在随机数字表中任选一数字并开始向某方向递推，遇到属于总体单位编号范围内的数字号码就确定该号码对应的单位为样本单位，一直到抽够规定的样本单位数为止。若是不重复抽样，则碰到重复的数字就舍去，并继续往下递推。

简单随机抽样有两种抽样的方法：重复抽样和不重复抽样。重复抽样是指从总体中抽取一个单位后，再把这个单位放回总体再抽取第二个单位，直至抽取第 n 个单位为止。这样，一个单位有可能被重复抽中，故称为重复抽样。不重复抽样是每次抽取一个单位后不再放回总体，而在所有未入样的单位中再进行等概率抽样，因此每个总体单位不可能被重复抽中，故称为不重复抽样。若总体中单位数目为 N，从中抽取 n 个单位为样本，从排列组合的方式看，采用重复抽样的方法共有 N^n 个可能的样本；而采用不重复抽样的方法则有 $C_N^n = \frac{N!}{n!(N-n)!}$ 个可能的样本。

简单随机抽样的特点是：从总体中抽取每一个单位时，各单位被抽中的机会均等，且方便简单，易于掌握。当我们要观察的总体中各单位之间的数据差异程度不大时，或其数据本身分布比较均匀，或我们对总体了解很少时，适合采用这种抽样方式。如果各单位之间相差较大或分布很不均匀时，不宜采用此种方式。

简单随机抽样是抽样调查的最基本的形式，本书在介绍统计推断方法时，除特别说明外，均适用于简单随机抽样方式。

三、分层抽样

分层抽样，又称类型抽样或分类抽样。这种抽样方法是先将总体各单位按某一主要标志分层(组)，而后在各层中按随机原则抽取若干个样本单位，由各层的样本单位组成一个样本。若总体中有单位数 N 个，将总体划分为 K 层，第 i 层总体单位数为 N_i(i =1, 2, 3, …, k)个，再从各层中随机抽取样本单位 n_i 个，构成样本。例如，我们研究的目标总体是某一人群时，可以将人群按性别、年龄等分组，再分别从男性、女性人群中各随机抽取一定比例的人员，或从不同年龄人群中再随机抽取具体的被调查者构成样本。再如，要调查电视节目的收视率，可以将目标总体划分为城市和乡村后，再分别从城市和乡村进行随机抽样；在农产量调查中，可按地形条件分为山区、丘陵、平原三层后再从各类中随机抽取样本单元等。当然，采用分层抽样需要事先对总体有一定的认识，或在获取有某些资料时可以对总体进行分层分类，同时还应该能够从各层各类中随机抽取一定的单位数。

分层抽样是将分组方法与随机抽样结合的抽样方法，是一种经常采用的方法，因为它具有如下优点。

(1)能够保证样本的代表性。因为在抽样前通过分层，可以把总体中某一标志特征比较接近的单位归为一层，将差异较大的分开，使各层内的各单位的分布比较均匀，而且每一层中的各单位都有被抽中的机会，使样本更接近于总体的分布，从而可以保证所抽取的样本具有一定的代表性。如按收入水平分组的分层抽样，可以保证收入高、中、低的人群均有单位进入样本。

(2)适当的分配各层样本量可以较大地提高抽样的精度。由于分层抽样是在各层中抽选

样本单位，按抽样推断的理论，内部差异越小，对总体的估计精度就越高。在各层中各单位之间的差异小于总体中所有单位之间的差异，影响总体估计精度的主要是各层的层内差异，因而在总体各单位之间的差异较大时，运用分层抽样可以得到比用简单随机抽样更准确的估计结果。

(3)分层抽样可以按自然的地区或行政系统分层，从而可以使抽样的组织与实施比较方便。此外，分层抽样调查除了可以用于估计总体的水平外，还可对每一层的水平进行估计。如在调查消费支出水平时，除了可以利用按收入分层的分层抽样的样本估计所有人的平均消费支出外，还可以对高、中、低等不同收入水平人群的消费支出分别进行估计，并进行比较分析。

采用分层抽样的方法时，在对总体分层以后，从各层抽取样本单位，可以根据各层的单位数等比例抽样，也可以是不等比例抽样。其中，等比例抽样就是从各层中按相同的比例抽取样本单位数，样本单位在各层的分配比例同总体单位在各层的分配比例相同。然而，当总体中各层的层内差异大小差别比较大时，为了提高样本的代表性和提高总体参数估计的精度和抽样效率，在抽样时，对内部差异较大的层抽取的样本比例高一些，对内部差异较小的层，其抽样比例可以低一些，这种方法被称为不等比例抽样。

四、等距抽样

等距抽样，又称机械抽样或系统抽样。它是将总体全部单位按某一标志特征排序，并在该序列中的前 N/n 个单元中随机抽取第一个样本单元，而后按固定的顺序和相等间隔在总体中抽取若干样本单位，构成样本。

等距抽样的最大优点在于这种抽样方法组织形式简便，易于实施。在设计这种抽样方案和抽取样本单位时，只要具备所调查总体的基本资料，如总体单位的标号、名单或某些方面的数据等，便可利用其构成总体抽样框，然后在此抽样框的基础上按相等的间隔抽取各样本单位。如进行农村经济抽样调查、人口抽样调查和产品质量抽样检验的过程中都广泛地采用了等距抽样。

在已知总体某些相关信息的情况下，采用等距抽样能保证样本单位在总体中均匀地分布，从而提高样本对总体的代表性，有利于提高对总体参数估计的精度。例如，在我国农产品产量抽样调查中，将总体单位按前三年或当年预计的粮食平均产量，由低到高顺序排序，这时总体各单位的粮食产量呈现线性变动趋势。按这种总体单位的顺序，等距地从总体抽取样本单位，这样的样本结构大致能反映出总体的结构，所以能取得较好的抽样估计效果。

等距抽样要计算抽样间隔。间隔 d 等于总体单位数 N 除以样本单位数 n，即 $d=\frac{N}{n}$。

例如，从 10 000 名职工中抽取 2%(即将 200 名)进行调查，抽样间隔为 50，职工可按姓氏笔画排列，然后在第 1～50 名职工中随机抽取第一人，如抽中的第一样本的顺序号是 5，则第二个样本单位顺序号则为 5+50=55 号，第三个样本单位的顺序号为 105，依此类推，最后一个样本单位的顺序号为 9955 号。

总体的排列顺序是针对总体单位的某一数据标志(即某一方面)而言的。当总体排序时所依据的数据标志跟所要调查的数据无关或基本无关时，这种排序称为无关标志排序。例如，调查职工生活水平时，将职工按其姓氏笔画排序，对产品质量检查，按产品入库顺序排序等都是无关标志排队。而当总体排序时所依据的数据标志就是所要调查的项目，或与所要调查

的项目有密切关系或有一定关系时，这种排序方法称为有关标志排序。例如，对农产量进行调查，把各地块按往年平均产量的高低排序，对职工家庭生活水平进行调查，按职工工资水平的高低进行排序等都是按有关标志排序。按有关标志排序的等距抽样其估计效果一般要好于无关标志排序的等距抽样。

等距抽样按样本单位抽选的方法不同，可分为随机起点等距抽样、半距起点等距抽样和对称等距抽样等。

(1)随机起点等距抽样。当抽样间隔 d 确定以后，在第一组(即前 d 个单位)随机抽选第一个样本单位，设顺序号为 a，则第二个样本单位的顺序号为 d+a，依此类推，第 n 个样本单位的顺序号为(n–1)d+a。当总体各单位按无关标志排序时，随机起点等距抽样是可以应用的。当总体各单位按有关标志排序时，随机起点等距抽样可能会产生系统性误差，即偏大或偏小。

(2)半距起点等距抽样。这种抽样方法要求各样本单位都选在各组的中点。各样本单位的顺序号是：第一个样本单位的顺序号是$\frac{d}{2}$，第二个样本单位的顺序号是$d+\frac{d}{2}$，……第 n 个样本单位的顺序号是$(n-1)d+\frac{d}{2}$。无论按有关标志排序还是按无关标志排序都可以采用这种方法。此法的优点是简单易懂，易于实践。当总体按有关标志排序时，中间项的数值的代表性比较强，从而保证了样本对总体有充分的代表性。当总体排序确定后，样本量确定，则样本单位也随之确定。

(3)对称等距抽样。这种抽样方法要求在前 d 个单位中随机抽取第一个样本单位，假设该单位的顺序号为 a，然后在第二组与第一个样本单位对称的位置抽取第二个样本单位，它的序号为 2d–a。在第三组与第二组样本单位对称的位置抽取第三个样本单位，它的序号是2d+a，以后抽出的样本单位序号依次为 4d–a, 4d+a, 6d–a, 6d+a,…，此方法保留了半距起点等距抽样的优点，又避免了它的局限性，使其优点更加明显。

五、整群抽样

整群抽样是将总体各单位划分为若干群，然后以群为单位，从总体中随机抽取一部分群，对被抽中群内的所有单位进行全面调查。整群抽样对总体划分群的基本要求是：第一，群与群之间不重叠，即总体中的任何一个单位只能属于某个群；第二，全部总体单位毫无遗漏，即总体内的任一单位必属于某个群。总体中各群内所包含的单位数可以是相同的，也可以不相同。

整群抽样划分群与分层抽样划分层的目的与要求有很大的区别：分层抽样时划分层的要求是将相近的总体单位划归同一层，从而减少层内单位的差异，分层抽样抽取的单位仍是总体的基本单位。例如，某城市进行居民住户调查，将该市居民户按某一标志特征划分为若干层，抽样调查的单位是居民户。而整群抽样中划分群的目的是扩大“总体单位”，抽取的单位不是总体的基本单位，而是总体的群。例如，调查城市居民住户，将该城市居民住户按居民委员会行政区域划分为若干群，抽样的单位是一个个群，即对抽中的样本群中的全部居民住户进行全面调查。

整群抽样的主要优点是设计和组织抽样比较方便，能节省人力、财力、物力和时间。整群抽样的缺点是，相对于简单随机抽样，在相同的调查单位下，用整群抽样估计总体时，估计精度相对较低。

六、多阶段抽样

当总体很大时，可把抽样过程分为几个阶段，到最后才能抽到具体的样本单位，这样的抽样形式被称为多阶段抽样。例如，二阶段抽样是将一个很大的总体划分为若干个样本群(称为一阶单位)，先从总体中抽取若干一阶单位，再从抽中的一阶单位中抽取若干个二阶单位进行调查观测和抽样推断。如果抽样是按更多个阶段进行，那么，可以继续在抽中的各二阶单位中抽取三阶单位，再从抽中的各三阶单位中抽取四阶单位……以最后阶段单位作为样本的基本单位，这就形成了多阶段抽样。

例如，对我国职工家庭收支情况调查，第一阶段先抽选调查城市；第二阶段从中选的城市中随机抽取调查行政单位；第三阶段再从抽中的行政单位中随机抽选职工，确定具体的调查户，调查其每月实际生活费收支情况。

多阶段抽样的主要优点是有利于抽样的组织和实施，可以提高抽样估计精度和满足各阶段对调查数据的需求。多阶段抽样还特别适用于大批量生产的产品检验，可节约人力、物力和财力，在实践中得到了广泛的应用。

第三节　总体、样本与抽样分布

一、总体分布

总体是一个具有确切分布的随机变量 X，总体分布就是 X 的分布。因此总体分布是指研究对象这一总体的各单位某变量的分布状况，即总体各单位数据分布状况的一种概括。研究对象不同，其总体分布不同。如新生婴儿的性别，或是男性或是女性，二者必居其一，在概率分布中属于两点分布。根据以往的大量统计资料可以看到，男女性别比例大致相同，男婴略多于女婴。又例如，人的身高一般近似地服从正态分布，接近平均身高的人较多，而特别高和特别矮的人都比较少，表现为中间多，两头少的分布。此外，统计资料表明，职工收入的分布一般属于正偏分布(如图 7-9 所示)，即中等及中等偏下收入的职工占大多数，而比较高和特别高收入的职工占少数。

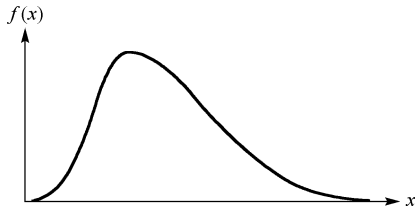


图 7-9　职工收入分布(正偏分布)

二、样本分布

如果我们能掌握总体分布就能掌握关于总体的一些重要信息，但由于总体的全部数据无法得到或没有必要花费高成本得到，因而可能无法掌握总体的分布。这样我们往往是从总体中随机抽取样本，利用样本获取信息从而去认识总体。当然，若从总体中抽取一个样本容量为 n 的样本，那么这个样本中所有单位的数据也是有差异的，并形成一个分布，称为样本分布。

由于样本是从总体中抽取的，因而其自然会表现出总体分布的一些特征；因为样本是从总体中随机抽取的，样本的随机性会导致每次抽取的样本虽然能够表现出总体的一些特征，但是每次抽取的样本之间不完全一致，也很难与总体分布完全一致，所以样本分布也称为经验分布。例如，某地区全部职工的收入服从正偏分布，如果从该地区全部职工中随机抽取人

数为 n 的职工作为样本，这 n 个职工的收入也存在一个分布，即样本分布。因为随机抽取的样本分布一方面要受总体分布的制约，另一方面受总体分布的影响，所以一般应与总体分布相近似。例如，由于高收入及特高收入的职工相对比较少，被抽中的概率相对较低，因此，随机抽中的职工中高收入者也会相对较少。但是另一方面，由于样本的抽取是随机的，因而样本分布不可能与总体分布完全一致，特别当抽取的单位数即样本容量比较小时，这种差别可能还比较大。当然随着样本容量的增大，样本的分布会逐渐接近总体分布。

三、抽样分布

我们按随机的原则从总体中抽取样本的目的是用样本统计量估计总体参数，如用样本平均数估计总体平均数、用样本比例估计总体比例、用样本方差估计总体方差等。但由于样本是随机的，因而样本平均数、比例、方差会随样本的不同而不同，即样本统计量是随机变量。为了用样本统计量估计总体参数，我们需要研究样本统计量围绕总体参数变动的规律性，研究样本统计量的分布，即抽样分布。

所谓抽样分布是指抽取的总体单位数即样本容量 n 一定时，从总体中按随机抽样的原则，所有可能抽取的样本统计量的分布，也称为样本统计量的概率分布。与样本分布是经验分布不同，抽样分布是一个理论分布。直观地看，如果总体单位数为 N，抽取样本的容量为 n，采用不重复抽样的方式，就有 C_N^n（不考虑顺序）或 A_N^n（考虑顺序）个可能的样本，每一个样本都可以计算一个样本的均值 $\bar{x}$、样本比例 p、样本方差 s^2 等统计量，这 C_N^n 个 $\bar{x}$、p 或 s^2 等统计量所形成的分布就是抽样分布。

例 7-8：为了介绍方便，我们假设一个总体中仅包括四个单位，其数值分别为 1、2、3、4。现从中随机抽取容量为 2 的简单随机样本，请列出样本均值 $\bar{x}$ 的分布，即 $\bar{x}$ 的抽样分布。

解：总体概率分布（如表 7-3 所示）。

表 7-3　概率分布表

X	1	2	3	4
P	0.25	0.25	0.25	0.25

在重复抽样下，共有 $4^2=16$ 个可能的样本。由于每个样本被抽中的概率相同，均为 1/16，所有可能的样本及均值如表 7-4 所示。

表 7-4　抽样组合表

样本单位 / 均值 / 样本单位	1	2	3	4
1	1	1.5	2	2.5
2	1.5	2	2.5	3
3	2	2.5	3	3.5
4	2.5	3	3.5	4

这样，样本均值的抽样分布如表 7-5 所示。

表 7-5　样本均值的抽样分布

均值	1	1.5	2	2.5	3	3.5	4
次数	1	2	3	4	3	2	1
概率	1/16	2/16	3/16	4/16	3/16	2/16	1/16

可见，总体分布表现为等概率分布，但其样本均值的抽样分布并不是等概率分布，即虽然样本分布与总体分布近似，但抽样分布与总体分布可能会有很大的差异。

四、中心极限定理

由于正态分布在统计学中具有特别重要的地位，因此关于寻找极限分布函数为正态分布的普遍条件，即在什么条件下随机变量将趋近于正态分布，就成为人们非常关注的问题。由此，学者研究并证明出了有关此问题的定理，即中心极限定理。

中心极限定理证明了随着样本容量 n 的不断增大，不论原来的总体是否是正态分布，其样本均值将会趋向于正态分布。在实际工作中运用抽样推断去认识研究对象时，其总体分布不一定是正态分布，但只要样本容量足够的大，其样本均值就趋向于正态分布，从而可以进行各种估计和检验，因而可以说，中心极限定理在抽样推断中起着十分重要的作用。

实际应用中，究竟样本容量多大才能使样本均值趋向于正态分布，这一方面取决于总体分布的形状和偏离正态分布的程度，另一方面取决于统计量的性质。大量的实践和模拟证明：随着 n 的增大，样本均值趋向于正态的速度是相当快的；当 $n \geqslant 30$ 时，均值就可以近似地服从正态分布。

第四节　常用的抽样分布

均值是一个描述总体数量特征的重要度量指标之一，要想知道总体的均值，最常用的方法是从总体中抽取样本并根据样本来推断总体均值。而要用样本统计量推断总体参数需要了解样本统计量如样本均值 $\bar{x}$ 的分布，这对于认识、把握总体均值是非常重要的。要确定一个分布需要明确两个方面的问题：一是弄清它的概率分布的形式；二是要了解这一分布的一些主要特征，本节主要介绍最常用、最基本的几个样本统计量的抽样分布。

一、样本均值 $\bar{x}$ 的抽样分布

(一) 样本均值 $\bar{x}$ 的抽样分布形式

关于样本均值 $\bar{x}$ 的抽样分布形式，这与原有总体分布及样本容量大小有关。

(1) 若原总体分布是正态分布，那么不论样本容量的大小，样本均值 $\bar{x}$ 的抽样分布都服从正态分布；

(2) 若原总体分布是非正态分布，当从总体中抽取的是一个大样本，即一般认为 $n \geqslant 30$ 时，由中心极限定理可知，其样本均值 $\bar{x}$ 的抽样分布也近似服从正态分布。当从总体中抽取的是小样本，则样本均值 $\bar{x}$ 的抽样分布不保证服从正态分布，不能按正态分布去推断总体的均值。

(二)样本均值 $\bar{x}$ 分布的主要数量特征

关于样本均值 $\bar{x}$ 分布的主要数量特征，我们主要讨论其数学期望值与方差。抽样分布的这两个数量特征值不仅和原总体分布的均值和方差有关，而且还与抽样方法是重复抽样还是不重复抽样有关。

(1)样本均值 $\bar{x}$ 分布的数学期望比较简单，可以证明无论是重复抽样还是不重复抽样，其数学期望始终等于总体均值，即 $E(\bar{x})=\mu$ ，其中 μ 是总体的均值。

(2)样本均值 $\bar{x}$ 分布的方差与抽样的方式有关。

第一，在重复抽样的条件下，样本均值 $\bar{x}$ 的分布的方差为总体方差(σ^2)的 $1/n$ ，即 $D(\bar{x})=\dfrac{\sigma^2}{n}$ 。

第二，若为不重复抽样，则 $D(\bar{x})=\dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$，其中 $\left(\dfrac{N-n}{N-1}\right)$ 称作有限总体不重复抽样的修正系数。

综上，在总体服从正态分布或大样本的条件下，样本均值服从正态分布，其抽样分布的特征值如表 7-6 所示。

表 7-6　抽样分布及特征值表

原总体均值和方差	抽样方式	$\bar{x}$ 的数学期望	$\bar{x}$ 的方差
有限总体 (μ,σ^2)	重复抽样	μ	σ^2/n
有限总体 (μ,σ^2)	不重复抽样	μ	$\dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right)$
无限总体 (μ,σ^2)	重复抽样	μ	σ^2/n
无限总体 (μ,σ^2)	不重复抽样	μ	σ^2/n

例 7-9：某企业从一批电子元件中随机抽取了 64 个元件，目的是了解该批元件的使用寿命。根据过去的经验，该企业生产的该种电子元件的使用寿命的标准差为 320 小时，要求计算样本平均寿命与总体均值相差 80 个小时以上的概率。

解：由于企业所生产的电子元件是大批量的，抽样比例虽然很小但已经是大样本，所以样本均值的抽样分布近似为正态分布，其均值为 μ，方差为 σ^2/n，根据题意，要求 $P(|\bar{x}-\mu|>80)$，计算时可以将其分解为两个部分，即：

(1) $P(\bar{x}>\mu+80)=P\left(\dfrac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}>\dfrac{\mu+80-\mu}{\sqrt{320^2/64}}\right)=P(Z>2)$；

(2) $P(\bar{x}<\mu-80)=P\left(\dfrac{\bar{x}-\mu}{\sqrt{\sigma^2/n}}<\dfrac{\mu-80-\mu}{\sqrt{320^2/64}}\right)=P(Z<-2)$。

经过计算得到：$P(|Z|>2)=0.0456$，即样本平均寿命与总体均值相差 80 小时以上的概率为 4.56%。

正态分布总体方差已知时的分布，设 X 服从正态分布，即 $X\sim N(\mu,\sigma^2)$，从中抽取容

量为 n 的样本，则样本均值 $\bar{x}$ 的抽样分布为 $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，通过标准化，$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ 服从标准正态分布。

当总体方差未知，用样本方差 s^2 估计总体方差时，$\bar{x} \sim N\left(\mu, \frac{s^2}{n}\right)$。

经过变换的样本统计量 $\frac{\bar{x}-\mu}{\sqrt{s^2/n}}$ 服从 t 分布，即 $t = \frac{\bar{x}-\mu}{\sqrt{s^2/n}} \sim t(n-1)$，称 t 服从自由度为 $n-1$ 的 t 分布。

t 分布与正态分布一样是对称分布，但一般情况下比标准正态分布平坦和分散。但当自由度增大时，t 分布也趋向于标准正态分布。因此在总体方差未知但抽取的是大样本时，即 n 较大时，$\frac{\bar{x}-\mu}{s/\sqrt{n}}$ 近似地服从标准正态分布。

t 分布广泛应用于正态总体方差未知且小样本时的估计和检验。

二、样本比例 p 的抽样分布

在实践中，有大量的需要掌握总体有关比例的问题，如产品的次品率、合格率、电视节目的收视率等。要了解总体的比例 P，往往也同样需要通过样本的比例 p 去推断，这样需要首先了解样本比例的抽样分布。

样本比例 p 是在 n 个样本单位中具有某种特征的单位所占的比例：

$$p = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad x_i = \begin{cases} 1 & \text{具有某种特征} \\ 0 & \text{不具有某种特征} \end{cases}$$

p 是一随机变量，随着 n 的增大，p 近似正态分布。

(1) 若采用重复抽样，其数学期望和方差分别为：

$$E(p) = P, \quad D(p) = \frac{P(1-P)}{n}$$

p 的抽样分布为：

$$p \sim N\left(P, \frac{P(1-P)}{n}\right)$$

式中，P 表示总体比例；p 表示样本比例。

(2) 若采用不重复抽样，其数学期望和方差分别为：

$$E(p) = P, \quad D(p) = \frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right)$$

p 的抽样分布为：

$$p \sim N\left(P, \frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right)\right)$$

例 7-10：某企业正常情况下生产产品的次品率为 8%，由于生产产品的批量较大，该企

业随机抽取 100 个产品进行检验，试求次品率在 7%～9%的概率。

解：因为 $n=100$ 属于大样本，故样本比例 p 近似服从正态分布。同时由于抽样比例较小，修正系数 $\frac{N-n}{N-1}$ 可以忽略。因此 p 的标准差为 $\sqrt{\frac{0.08\times0.92}{100}}=0.0271$。

这样：

$$P(0.07<p<0.09)=P\left(\frac{0.07-0.08}{0.0271}<Z<\frac{0.09-0.08}{0.0271}\right)=P(-0.37<Z<0.37)=0.289$$

故产品次品率在 7%～9%之间的概率为 28.9%。

三、样本方差的抽样分布

所谓样本方差的抽样分布，是指抽取样本容量为 n 的样本时，由所有可能的样本的方差取值所形成的频率分布。

设 $X\sim N(\mu,\sigma^2)$，则 $Z=\frac{\bar{x}-\mu}{\sigma}\sim N(0,1)$，令 $Y=Z^2$，则 Y 为自由度为 1 的 χ^2 分布，记为 $Y\sim\chi^2(1)$。

进一步导出：当总体 $X\sim N(\mu,\sigma^2)$，从中抽取样本容量为 n 的样本，则

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2}\sim\chi^2(n-1)$$

将上式做一变换即可得到：

$$\frac{(n-1)s^2}{\sigma^2}\sim\chi^2(n-1)$$

χ^2 分布通常可用于总体方差的估计和许多非参数检验。χ^2 分布具有如下性质和特点。

(1) χ^2 分布的变量值始终为正值。

(2) χ^2 分布的形状取决于其自由度 n 的大小，通常 χ^2 分布为不对称的正偏分布，但随着自由度的增大逐渐趋于对称。

(3) χ^2 分布的期望值为 $E(\chi^2)=n$，方差 $D(\chi^2)=2n$。

(4) χ^2 分布具有可加性。若 U 和 V 为两个独立的 χ^2 分布随机变量，$U\sim\chi^2(n_1)$, $V\sim\chi^2(n_2)$，则随机变量 $U+V$ 服从自由度为 n_1+n_2 的 χ^2 分布。

四、两个样本均值之差的抽样分布

为推断两个总体的均值之差，我们需要独立地从两个总体中分别抽取样本容量为 n_1 和 n_2 的样本。在重复选取样本容量 n_1 和 n_2 的样本时，由两个样本均值之差的所有可能取值形成的频率分布，称为两个样本均值之差的抽样分布。

假定独立地从两个总体中抽取样本。其中，从总体 1 中抽取样本容量为 n_1 的样本，其样本均值为 $\bar{x}_1$，从总体 2 中抽取样本容量为 n_2 的样本，其样本均值为 $\bar{x}_2$，当两个总体都为正态分布时，两个样本均值之差 $\bar{x}_1-\bar{x}_2$ 的抽样分布为正态分布，此分布的数学期望为两个总体均值之差，即 $E(\bar{x}_1-\bar{x}_2)=\mu_1-\mu_2$，其分布的方差 $\sigma^2_{\bar{x}_1-\bar{x}_2}$ 为各自方差之和，$\sigma^2_{\bar{x}_1-\bar{x}_2}=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$。

$\bar{x}_1-\bar{x}_2$的抽样分布可表示为：

$$\bar{x}_1-\bar{x}_2 \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

当总体为非正态分布，而n_1和n_2比较大（一般要求$n_1 \geqslant 30, n_2 \geqslant 30$）时，两个样本均值之差的抽样分布近似服从正态分布。

当总体为正态分布，总体方差σ_1^2、σ_2^2未知时，两个样本均值之差的抽样分布服从t分布，但若当n_1和n_2比较大（一般要求$n_1 \geqslant 30, n_2 \geqslant 30$）时，其分布近似服从正态分布。

五、两个样本比例之差的抽样分布

所谓两个样本比例之差的抽样分布，是指从两个服从二项分布的总体中，分别独立地抽取容量为n_1和n_2的样本，在重复选取容量为n_1和n_2的样本时，由两个比例之差的所有可能取值形成的概率分布。

当两个样本均为大样本时，两个样本比例之差p_1-p_2的抽样分布近似地服从正态分布，其数学期望为$E(p_1-p_2)=P_1-P_2$，其分布的方差$\sigma_{p_1-p_2}^2$为各自方差之和，$\sigma_{p_1-p_2}^2=\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}$，大样本时的$p_1-p_2$的抽样分布可表示为：

$$p_1-p_2 \sim N(P_1-P_2, \frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2})$$

六、两个样本方差比的抽样分布

所谓两个样本方差比的抽样分布从两个正态分布的总体中分别独立地抽取样本容量为n_1和n_2的样本，在重复抽选容量为n_1和n_2的样本时，由两个样本方差比的所有可能取值形成的概率分布。

在抽样中，设总体$X_1 \sim N(\mu_1, {\sigma_1}^2)$，$X_2 \sim N(\mu_2, {\sigma_2}^2)$，分别从中抽取容量为$n_1$和$n_2$的样本。则：

$$\frac{\sum_{i=1}^{n_1}(x_{1i}-\bar{x}_1)^2}{{\sigma_1}^2}=\frac{(n_1-1)s_1^2}{{\sigma_1}^2} \sim \chi^2(n_1-1)$$

$$\frac{\sum_{i=1}^{n_2}(x_{2i}-\bar{x}_2)^2}{{\sigma_2}^2}=\frac{(n_2-1)s_2^2}{{\sigma_2}^2} \sim \chi^2(n_2-1)$$

两个独立的χ^2分布分别除以自由度（n_1-1, n_2-1）后相比得到的统计量服从F分布，即：

$$\frac{\frac{(n_1-1)s_1^2}{\sigma_1^2(n_1-1)}}{\frac{(n_2-1)s_2^2}{\sigma_2^2(n_2-1)}}=\frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1, n_2-1)$$

F分布广泛应用于方差分析、回归分析和协方差分析等。

思考与练习

1．举例说明什么是总体分布、样本分布和抽样分布。

2．解释中心极限定理的含义及其应用意义。

3．分析样本统计量分布与总体分布的关系。

4．从均值为 500、方差为 225 的总体中，利用简单随机抽样的方法抽取一个样本容量为 100 的样本，其样本均值服从什么分布？样本均值的期望值和方差值各是什么？

5．举例说明随机抽样中的各种抽样方法的具体做法及其特点与应用条件。

6．请说明二项分布与正态分布之间的关系。

7．设离散随机变量 X 的分布如下表所示：

概率分布表

X	−1	0	1
p	0.25	0.5	0.25

求：(1)变量 X 的期望值；

(2)变量 X 的方差。

8．已知一批产品的的重量服从正态分布，其平均重量为 100 克，标准差为 9 克，试问：

(1)从该批产品中随机抽取 1 件，其重量不到 95 克的可能性是多大？

(2)从该批产品中随机抽取 36 件，其平均重量不到 95 克的可能是多大？

第八章　参 数 估 计

第一节　样本估计量

一、总体参数与样本估计量

统计研究的目的是通过数据认识总体，从而得出相关的结论。在掌握总体全部数据的条件下，我们可以利用描述统计的方法进行分析与研究。但是，现实中我们的常态是无法掌握或没有必要掌握总体数据，这时就需要利用抽样推断的方法去认识总体现象。

抽样推断就是从研究对象的全部单元中随机抽取一部分单元进行调查并取得数据，并从这些数据中获取信息，以此来推断总体的数量特征。抽样推断最主要的内容包括参数估计与假设检验。

所谓参数是指描述总体数量特征的量。从狭义上讲是指决定某一理论分布的分布函数中某一个或若干个数值，如正态分布的期望值 μ 和方差 σ^2，它决定了变量分布的重心位置和离散程度；从广义上讲，参数是反映总体数量特征和决定有关模型的数值，除总体的期望值和方差外，还包括反映变量之间数量依存关系的回归模型的回归系数，相关模型中的相关系数等内容。统计是研究客观现象的数量表现及其规律性的，因此其主要目的就是要取得各种参数。当然，了解参数的方法除了全面调查外，还可以通过抽样调查利用样本数据加以估计。如用所抽取的部分职工的工资计算平均数去估计全部职工的平均工资，用产品的抽样合格率估计全部产品的合格率等。

总体是客观现象或数据的全体，只要总体的范围一旦确定，参数就会确定，它是一个不变的常数。在对总体参数进行估计时需要利用样本的统计量，这些统计量的具体值是根据样本数据计算得到的，由于样本的随机性，决定了样本统计量是一个随机变量。这些用于估计总体参数的样本统计量即样本函数的名称，称为估计量，如样本算数平均数是估计总体均值的估计量。而抽取一个样本并计算这个样本估计量所得到的数值称为估计值，它是估计量的具体表现。

二、评价估计量的标准

估计总体参数时需要设计相应的估计量，如对于总体均值可以用样本算术平均数，也可以用中位数、众数和几何平均数等来进行估计。然而在估计时，我们并不知道总体数据的均值，那么该如何找到最接近总体数据均值的估计量呢？在对总体参数进行估计时，需要选择一个好的估计量。选择估计量，需要比较不同估计量的抽样分布，因为抽样分布比较全面地概括了统计量的所有可能结果。性质优良的估计量一般要满足以下几个方面的要求。

第一，优良的估计量是随机变量，即使它的取值会随着样本的随机性而随机波动，但总体上应集中在参数真值的附近，围绕其对称地变化，即估计量应不存在系统偏差。

第二，估计量数值的随机变化程度小，估计的效果比较好。

第三，当样本容量不断增大时，估计量的值要能稳步地趋向总体参数的真值。

第四，样本数据来自于总体数据，它当然包含了总体的信息，这些信息构成了样本推断总体参数的依据，为了提高估计的精度，估计量要能充分地吸收样本中所包含的关于参数的信息。

第五，当假定的理论分布与总体数据的实际分布存在差距时，估计量的估计值不应受到太大的影响。

与上面的要求相对应，统计学家们在评价估计量时总结出五个评价标准：无偏性、有效性、一致性、充分性和稳健性。在实际应用中，比较强调的是前三个性质，下面逐一介绍。

(一) 无偏性

无偏性是指估计量抽样分布的数学期望等于总体参数的真值。假定总体有 N 个单位，按不重复、不考虑顺序的方法随机抽取 n 个单位组成样本，则样本总数为 C_N^n，对每一个样本观察值，都可以计算一个统计量的值，如果 C_N^n 个估计值的平均数等于参数的真值，则称该统计量是无偏的。

设总体参数为 θ，所选择的估计量为 $\hat{\theta}$，如果 $E(\hat{\theta})=\theta$，则称 $\hat{\theta}$ 为 θ 的无偏估计量。

如图 8-1 所示，左图中的估计量 $\hat{\theta}$ 的期望值为总体真值 θ，即 $E(\hat{\theta})=\theta$，为无偏估计量。而右图中的统计量 $\hat{\theta}$ 的期望值与总体真值不等，即 $E(\hat{\theta})\neq\theta$，为有偏估计量。

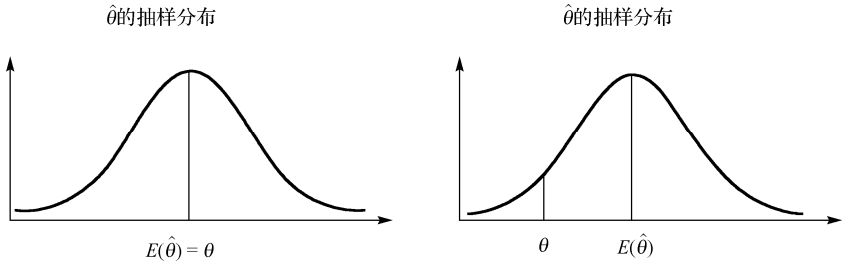


图 8-1　无偏估计量(左)和有偏估计量(右)

(二) 有效性

有效性是指估计量的方差尽可能小。一个无偏的估计量并不意味着一个样本的估计值就非常接近被估计的总体参数的真值。估计量的值与参数真值接近的程度是用估计量的方差来度量的。同一总体参数的两个无偏估计量相比，有更小方差的估计量为更有效的估计量。

由图 8-2 可以看到，$\hat{\theta}_1$ 与 $\hat{\theta}_2$ 均是 θ 的无偏估计量，但 $\hat{\theta}_1$ 的方差比 $\hat{\theta}_2$ 的方差小，说明估计量 $\hat{\theta}_1$ 有更多的估计值将会落在总体参数真值的附近，即 $\hat{\theta}_1$ 比 $\hat{\theta}_2$ 更有效。

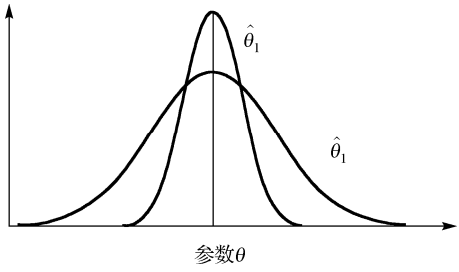


图 8-2　估计量 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 的抽样分布

(三)一致性

一致性是指随着样本容量的增大，即随着抽取的样本单元数的增多，估计量的估计值越来越接近于总体参数的真值。如果一个估计量是一致估计量，那么样本越大就越精确，因而可以通过增加样本容量来提高参数估计的精度和增加估计的可靠性。如果不是一致估计量，即使抽取大样本也无法提高估计的精度与可靠度，其结果就是浪费时间和费用。

第二节　区间估计的基本思想

一、点估计与区间估计

参数估计有点估计和区间估计两种。

(一)点估计

点估计就是用估计量 $\hat{\theta}$ 的某个取值直接作为总体参数的估计值。如用随机抽取的部分职工的平均工资 $\bar{x}$ 直接作为全部职工平均工资 μ 的估计值，用抽取的部分学生的考试及格率 p 直接作为全部学生考试的及格率 P，用抽取的部分产品的合格率直接作为全部产品合格率的估计值等。

由于样本是随机抽取的，一个被抽中的具体的样本得到的估计值很可能不同于另一个(或另一次)样本的估计值，也很可能不同于总体参数的真值。点估计的缺陷是没有办法给出参数估计的可靠性，也无法确定点估计值与总体参数真值接近的程度，因为一个点估计量的可靠性是由其抽样分布的标准误差来衡量的。因此，我们对总体参数真值进行估计时，往往不能完全依赖一个点估计值，而应围绕点估计值构造出总体参数的一个区间。

(二)区间估计

(1)区间估计是在点估计的基础上给出总体参数估计的一个区间，该区间通常是由样本估计量的值即点估计值加减估计误差得到的。例如，从某批产品中随机抽取一个容量为 n 的样本，并计算得出样本的平均使用寿命为 1000 小时，则该批产品平均寿命参数的这一点估计值为 1000 小时。但实践中仅靠这样一个点估计值往往是不够的，全部产品平均寿命正好为 1000 小时的可能性很低。因此，对总体参数的估计往往需要给出一个可控制的范围，即提出该批产品平均使用寿命的上限和下限，如 800～1200 小时之间，这就是估计的区间。当然，如果只给出参数的可能取值范围，并没有明确指出参数究竟会取哪一个值，从这一点看，估计的区间似乎没有点估计那样清晰。但是，区间估计除了给出估计的区间外，还需要说明估计结果的置信度，并能把估计的置信度与估计的区间有机地联系起来，这就是区间估计的好处。

从上面的描述可以看到，区间估计是指根据样本资料给总体参数划出一个大致的范围，以期该范围能覆盖参数的真实值，在给出这一范围的同时，给出相应的概率作为估计置信度的一个度量。

区间估计是用样本计算的估计量的值去估计总体参数，由于估计量是随机变量，那么由估计量构成的区间也应是随机区间。既然是随机区间，它可能包含了总体参数，也可能不包含总体参数，那么做出的区间包含了总体参数的可能性有多大呢？这种把估计区间与置信度联系起来的区间，称为置信区间。

(2)置信度。设 $x_1,x_2,\cdots,x_n$ 为总体的一个样本，θ 为总体参数，由样本确定的估计量为 $\hat{\theta}_1=\hat{\theta}_1(x_1,x_2,\cdots,x_n)$ 和 $\hat{\theta}_2=\hat{\theta}_2(x_1,x_2,\cdots,x_n)$，对于给定的 $\alpha(0<\alpha<1)$，如果使 $P(\hat{\theta}_1\leqslant\theta\leqslant\hat{\theta}_2)=1-\alpha$ 成立，则称 $[\hat{\theta}_1,\hat{\theta}_2]$ 为 θ 的置信区间。其中 $\hat{\theta}_1$ 为置信区间的下限，$\hat{\theta}_2$ 为置信区间的上限，$1-\alpha$ 则被称为置信度或置信水平。

从统计学的角度来说，总体参数是确定的值，不存在是否会落在某个区间的问题，相反倒是存在着建立的估计区间能否包含着参数的真实值的问题。因为用样本构造的估计区间是不固定的，是一个随机区间，会随着样本的不同而变化。在利用样本数据对总体参数进行抽样推断的实践中，我们往往只抽取一个样本，而利用此样本所构造的区间是与该样本相联系的、在一定的置信度下的置信区间，这一区间有可能包含总体参数，也有可能不包含。置信度可以直观地理解为：如果抽取 100 个样本，可以构造 100 个置信区间，这样将会有 $100(1-\alpha)$ 左右个区间包含总体参数的真值，而不包含参数真值的将有 100α 左右个区间。

如图 8-3 所示，图 8-3 中的中间横线表示总体真值，若从某总体中随机抽取 20 个样本，将会得到 20 个由样本所构造的置信区间(置信度为 95%)，这之中将有 19 个区间包含了总体参数(均值)的真值，一个(即 11 号样本)区间没有包含总体参数真值。当然，进行区间估计时，我们一般只抽取一次(个)样本，不会抽取很多次，这时只能希望所构造的区间是大量包含总体参数真值区间的一个，但又无法避免它也有可能是少数几个不包含参数真值的样本区间中的一个。

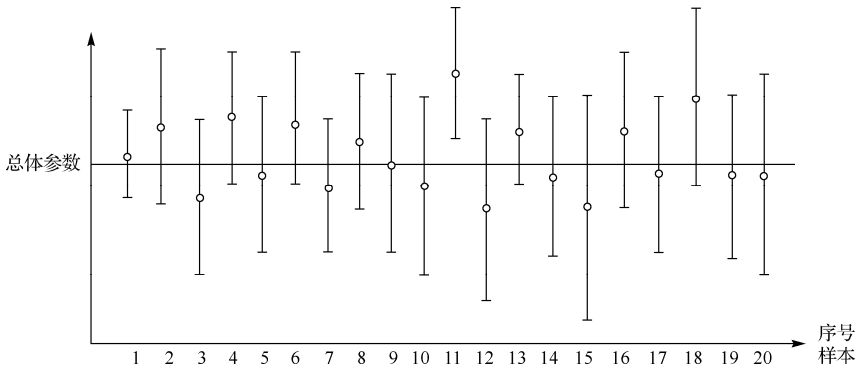


图 8-3　参数估计的 20 个置信区间

二、区间估计的基本思想

对总体参数进行区间估计，就是在已知样本统计量抽样分布的基础上，利用样本统计量估计总体参数。下面以总体均值的区间估计为例，说明区间估计的基本思想。

现已知(或假设)总体服从正态分布，即 $X\sim N(\mu,\sigma^2)$，若从总体中随机抽取一个容量为 n 的样本，则样本均值 $\bar{x}$ 的抽样分布为 $\bar{x}\sim N\left(\mu,\dfrac{\sigma^2}{n}\right)$，对其进行标准化，得到服从正态分布的统计量 Z：

$$Z=\frac{\bar{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}}\sim N(0,1)$$

那么，在 1–α 的置信度下，有：

$$P\left(\left|\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| \leqslant Z_{\alpha/2}\right)=1-\alpha$$

或

$$P\left(-Z_{\alpha/2} \leqslant \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leqslant Z_{\alpha/2}\right)=1-\alpha$$

上式也可以写成下面形式：

$$P\left(\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)=1-\alpha$$

即在 1–α 的置信度下，总体参数的区间为：

$$\left(\bar{x}-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{x}+Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

从上式可以得出：

当置信度$1-\alpha$为 68.27%时，置信区间为$\left(\bar{x}-1\frac{\sigma}{\sqrt{n}},\bar{x}+1\frac{\sigma}{\sqrt{n}}\right)$；

当置信度$1-\alpha$为 95%时，置信区间为$\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}},\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$；

当置信度$1-\alpha$为 95.45%时，置信区间为$\left(\bar{x}-2\frac{\sigma}{\sqrt{n}},\bar{x}+2\frac{\sigma}{\sqrt{n}}\right)$；

当置信度$1-\alpha$为 99.73%时，置信区间为$\left(\bar{x}-3\frac{\sigma}{\sqrt{n}},\bar{x}+3\frac{\sigma}{\sqrt{n}}\right)$；

从置信区间可以看到，置信区间越大，估计结果即估计误差就越大，若置信区间过大，估计也就失去了意义。置信区间的大小与置信度的高低存在着正向变动关系，要提高估计的置信度，当然估计区间越大越好，但置信区间的大小与估计的精确程度呈反向关系。因此，在样本大小一定的情况下，既要提高估计精度，降低估计的误差，又要保证较高的置信度，是不容易做到的。精度和置信度在区间估计中相互矛盾，优良的区间估计，不能仅偏重于某一方面的要求，而要把精度和置信度兼顾起来。

第三节　一个总体参数的区间估计

虽然区间估计的原理相同，但当所具备的条件不同(如总体分布是否已知、总体方差是否已知、随机抽取的是大样本或小样本、抽样方式是重复抽样或不重复抽样等)时，总体参数估计的具体做法还存在差异。

一、总体均值的区间估计

已知 $x_1, x_2, \cdots, x_n$ 为来自均值为 μ（未知）、方差为 σ^2 的总体的一个随机样本，样本均值为 $\bar{x} = \frac{1}{n}\sum x_i$，要求在 $1-\alpha$ 的置信度下，对 μ 进行区间估计。

（一）从总体方差 σ^2 已知的正态分布总体中随机抽取样本时总体均值的置信区间

如果总体服从正态分布，且总体方差 σ^2 已知，则可直使用上节区间估计的公式。

置信度 $1-\alpha$ 条件下总体均值的置信区间为：

$$\left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

其中，$Z_{\alpha/2}$ 与给定的置信度有关，可以通过查正态分布表得到。

一般，总体均值的区间估计有如下几个步骤。

第一，确定置信度，置信度的确定可以根据估计的需要确定，如 90%、95%、99%。

第二，根据置信度确定 $Z_{\alpha/2}$ 值。

第三，抽取一个容量为 n 的样本。

第四，计算样本平均数 $\bar{x}$ 和样本平均数的标准差 $\sigma_{\bar{x}}$。重复抽样时样本平均数的标准差 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$，有限总体、不重复抽样时的标准差 $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$。

第五，构造置信区间 $\left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$。

例 8-1：某企业生产一批电子元件，为了解该批产品的平均耐用时间，从该批产品中采用简单随机抽样的方法随机抽取了 250 个元件，测量并计算出样本的平均耐用时间为 6500 小时。根据以往资料得知该企业生产的电子元件服从正态分布，总体标准差为 150 个小时，现要求以 95%的置信度构造该批电子元件平均耐用时间的置信区间。

解：本题中，已知总体数据服从正态分布，总体方差已知，置信度为 95%时 $Z_{\alpha/2} = 1.96$。所以，总体平均数的置信区间为：

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6500 \pm 1.96 \times \frac{150}{\sqrt{250}} = 6500 \pm 18.6$$

即在 6481.4～6518.6 个小时之间。

（二）从已知方差 σ^2 的非正态分布总体中随机抽取一大样本时的总体均值置信区间

在很多情况下，虽然我们要估计的是非正态分布总体的参数，但根据中心极限定理，当样本容量 n 足够大时，无论总体服从什么分布，$\bar{x}$ 的抽样分布将近似地服从正态分布。因此这时我们可以用 $\left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ 来近似地估计出总体均值 μ 的置信区间。

例如，上例中是已知生产的电子元件服从正态分布，但若事先没有关于总体分布的信息，不知道该批电子元件总体上服从什么分布，但由于抽取了 n=250 的大样本，又根据历史数据了解到总体数据标准差大概为 150 个小时，这时区间估计的方法与上面的情况相同，即：

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 6500 \pm 1.96 \times \frac{150}{\sqrt{250}} = 6500 \pm 18.6$$

总体平均数在95%的置信度下的区间在6481.4～6518.6小时之间。

(三)从非正态分布总体中抽取一大样本，且总体方差σ^2未知时的总体均值的置信区间

若随机抽取一大样本，不论其是来自何种分布的总体，根据中心极限定理样本均值的抽样分布近似地服从正态分布：$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$。

这样$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$，此时，可利用样本均值估计总体均值置信区间的公式为：

$$\left(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

若总体方差σ^2未知，则可以用样本的方差s^2作为总体方差的估计量。这时，$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$则服从自由度为$n-1$的$t$分布。随着自由度的增大，$t$分布也趋向于标准正态分布，因而总体均值的置信区间可用下面公式计算：

$$\left(\bar{x} - Z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

(四)从正态分布总体中随机抽取一样本，且总体方差σ^2未知时的总体均值的置信区间

设X服从正态分布，即$X \sim N(\mu, \sigma^2)$，从中抽取容量为n的样本，则样本均值$\bar{x}$服从正态分布，即$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$，则$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$。但当总体方差未知时，用样本标准差$s$作为总体标准差$\sigma$的估计值，则$\frac{\bar{x}-\mu}{s/\sqrt{n}}$服从自由度为$n-1$的$t$分布。

但是，如上面所述随着自由度的增大，t分布也趋向于标准正态分布。因此在总体方差未知但是大样本时，即n较大时，$\frac{\bar{x}-\mu}{s/\sqrt{n}}$将近似地服从标准正态分布。这样，我们可以得到总体均值的置信区间的公式。

(1)大样本下，置信度$1-\alpha$下的区间为：

$$\left(\bar{x} - Z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + Z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

(2)小样本下，置信度$1-\alpha$下的区间为：

$$\left(\bar{x} - t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$$

例 8-2：某电子元件厂要估计新型产品的平均使用寿命，在生产线上随机抽取了 9 个元件并进行了测试，取得了以下数据(小时)：

5100　5100　5400　5260　5400　5100　5320　5180　4940

假设该生产线上生产的产品的使用寿命服从正态分布，要求在 95%的置信度下估计这种新产品的平均使用寿命的置信区间。

解：根据已知信息得知，产品使用寿命服从正态分布，总体方差 σ^2 未知，$n=9$ 为小样本，经计算可得：样本均值 $\bar{x}$=5200，样本标准差 s=156.5。95%置信度下，$t_{\alpha/2}(9-1)=2.306$。

则置信区间为：$\bar{x}\pm t_{\alpha/2}\frac{s}{\sqrt{n}}=5200\pm 2.306\times\frac{156.5}{\sqrt{9}}$。

这种新产品的平均使用寿命的置信区间为 5079～5320 个小时之间。

二、总体比例的区间估计

总体比例是指总体中具有某种特征的单元所占的比例，也是实际工作中常见的要估计的参数之一。例如，要估计产品的合格率、对公共服务表示满意的人数占全部人员的比例、一片林木中遭受病害的比例、全部企业中亏损企业所占的比例等。它实际上是总体均值的一种特例。

当反映每个样本单元特征的变量是一个定性变量(记为 X)，在其具有某种特征(如产品为合格)时用 1 表示，不具有该种特征(如产品为不合格)时用 0 表示，若总体中包含有 N 个单元，则具有某种特征所占的比例(如合格率) $P=\frac{1}{N}\sum_{i=1}^{N}x_i$，若从中随机抽取 n 个单元作为样本，则样本比例 $p=\frac{1}{n}\sum_{i=1}^{n}x_i$。样本比例是总体比例的一个无偏估计量，可以根据样本比例的抽样分布，利用样本比例进行总体比例的区间估计。

对总体比例进行估计时，需要用一个随机的大样本。而当样本量足够大时，样本比例 p 近似地服从正态分布，期望值为 $E(p)=P$，方差为 $D(p)=\frac{PQ}{n}$，其中($Q=1-P$)，由于实际工作中总体比例 P 是未知的，我们所要估计的也正是这个总体比例 P，所以用区间估计时就需要用样本比例 p 来估计 $D(p)$。

如果 nP 和 $n(1-P)$ 两者皆大于 5，并且 n 相对总体来说很小，则在置信度 $1-\alpha$ 下的总体比例的区间估计公式为：

$$p\pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

如果研究总体是有限的，尤其是抽样比例较大时，就要采用有限总体修正系数，从而总体比例的区间估计公式为：

$$p\pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

例 8-3：某企业产品质检部门为检查企业生产的产品的合格率，在其所生产的产品中随机抽取了 120 个产品进行检验，检验结果为 120 个产品中有 114 个合格产品，计算得到样本

合格率为 95%。现要求在 95%的置信度下估计该企业产品合格率的置信区间。

解：已知 $n=120$，$p=\dfrac{114}{120}=95\%$，根据置信区间公式得：

$$p\pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}=0.95\pm1.96\times\sqrt{\frac{0.95(1-0.95)}{120}}=95\%\pm1.99\%$$

即(93.01%，96.99%)，估计出的该企业产品合格率的置信区间为 93.01%～96.99%。

三、总体方差的区间估计

现实中，我们有时需要对反映数据之间差异程度或数据偏离总体平均水平程度的方差进行估计。如企业生产的电子元件平均寿命虽然合乎要求，但若电子元件使用寿命的方差很大，那么这些产品的质量还是有问题的，因而我们在了解其平均寿命的同时，还往往需要了解其总体方差或标准差的大小。

总体方差 σ^2 通常是未知的，需要我们通过样本数据对其进估计。在总体正态分布的条件下，统计量 $\dfrac{(n-1)s^2}{\sigma^2}$ 近似地服从 $n-1$ 的 χ^2 分布，记作 $\chi^2(n-1)$。在 $1-\alpha$ 的置信度下：

$$\chi^2_{1-\alpha/2}<\frac{(n-1)s^2}{\sigma^2}<\chi^2_{\alpha/2}$$

对上式通过不等式的变换，可得到 σ^2 的置信区间如下：

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}}<\sigma^2<\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

例 8-4：电子元件厂生产的某电子元件的使用寿命服从正态分布，为了检验产品生产的稳定性，企业质检部门从生产的产品中随机抽取了 31 个产品并进行了测试，根据样本数据计算出的样本产品的使用寿命的方差为 300 小时。现希望以样本数据为基础对该种电子元件使用寿命的方差 σ^2 进行区间估计，要求置信度为 95%。

解：根据已有的信息得知，总体分布近似地服从正态分布，利用样本数据得到的样本方差 $s^2=300$，$n-1=31-1=30$，置信度为 95%，查自由度为 30 的 χ^2 分布表得到 $\chi^2_{\alpha/2}=\chi^2_{0.025}(30)=46.979$，$\chi^2_{1-\alpha/2}=\chi^2_{0.975}(30)=16.791$。

代入公式得：

$$\frac{(31-1)\times300}{46.979}<\sigma^2<\frac{(31-1)\times300}{16.791}$$

$$191.6<\sigma^2<536$$

即 σ^2 的 95%的置信区间为 191.6～536。

第四节　两个总体参数的区间估计

在实际生活与工作中，我们经常要对来自两个不同总体的参数进行比较，如比较两个地区职工的平均工资、两个单位的产品合格率或是两个品牌产品的使用寿命等。但是比较时，我们往往无法直接得到总体的数据，因而只能通过样本数据进行估计。

一、两个总体均值差的区间估计

设 $x_1,x_2,\cdots,x_n$ 为来自均值为 μ_X、方差为 $\sigma^2{}_X$ 的总体的随机样本，$y_1,y_2,\cdots,y_m$ 为来自均值为 μ_Y、方差为 $\sigma^2{}_Y$ 的总体的随机样本，且相互独立，$\bar{x}$、$\bar{y}$ 分别为两个样本的平均值，s_x^2、x_y^2 为样本方差。要估计 X 和 Y 两个总体的均值之差的置信区间，需要在两个样本均值之差的基础上，利用样本均值之差的抽样分布。

(一)两个总体服从正态分布(或非正态分布，大样本)且方差已知

(1)若两个总体均服从正态分布，则从其中随机抽取的样本容量分别为 n 和 m 两个独立的样本的样本均值 $\bar{x}$、$\bar{y}$ 依然服从正态分布，因而 $\bar{x}-\bar{y}$ 也服从正态分布。

当 $X\sim N(\mu_X,\sigma_X^2)$，$Y\sim N(\mu_Y,\sigma_Y^2)$ 时，

$$\bar{x}-\bar{y}\sim N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right)$$

这样，

$$\frac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}\sim N(0,1)$$

由此可以得到在 $1-\alpha$ 置信度下的两总体均值之差 $\mu_X-\mu_Y$ 的置信区间为：

$$(\bar{x}-\bar{y})\pm Z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

(2)两个总体也许不服从正态分布，但只要从两个总体中随机抽取样本容量足够大的样本，那么根据中心极限定理，样本均值的抽样分布近似服从正态分布，即：

$$\bar{x}\sim N\left(\mu_X,\frac{\sigma_X^2}{n}\right),\quad \bar{y}\sim N\left(\mu_Y,\frac{\sigma_Y^2}{m}\right)$$

两个样本相互独立时，样本平均数之差的抽样分布为：

$$\bar{x}-\bar{y}\sim N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right)$$

进而得到一定置信度下两个总体均值之差的区间估计：

$$(\bar{x}-\bar{y})\pm Z_{\alpha/2}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$$

例 8-5：从两地各随机抽取 25 户居民家庭调查其子女一个月的课外教育费用支出，利用样本数据计算得到：A 地平均消费支出 450 元，B 地平均消费支出 325 元。已知两个地区居民子女课外教育费用支出均服从正态分布，且 A 地总体方差为 750，B 地总体方差为 850，要求在 95%置信度下构造两地区居民家庭每月在子女课外教育上的平均支出差异的置信区间。

解：由于两总体均服从正态分布，因此 $\bar{x}_A-x_B$ 服从正态分布，根据上面区间估计公式得到以下公式：

$$(\bar{x}_A-\bar{x}_B)\pm Z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n}+\frac{\sigma_B^2}{m}}=(450-325)\pm 1.96\times\sqrt{\frac{750}{25}+\frac{850}{25}}=(109.32,\ 140.68)$$

即两地区居民家庭每月在子女课外教育费用上的平均支出之差的 95%的置信区间是 109.32～140.68 元之间。

（二）两个总体服从正态分布，小样本且总体方差未知但相等

如果两个总体均服从正态分布，并从这两个总体中各随机抽取一个独立的随机小样本时，如前所述，两个样本均值之差服从正态分布：

$$\bar{x}-\bar{y}\sim N\left(\mu_X-\mu_Y,\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right)$$

当总体方差未知时，需要用样本数据估计总体方差。又由于已知两总体方差相等，因而可以利用样本数据估计一个共同的样本方差 s_p^2。其计算方法是首先要利用样本数据分别计算两个样本的方差 s_x^2、s_y^2，然后计算两个样本方差的加权平均数，权数则是它们的自由度，即：

$$s_p^2=\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}$$

这时，$\bar{x}-\bar{y}$ 的方差为 $\frac{s_p^2}{n}+\frac{s_p^2}{m}=s_p^2\left(\frac{1}{n}+\frac{1}{m}\right)$。

可以证明：$t=\dfrac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ 服从自由度为 $n+m-2$ 的 t 分布。这样可以估计 $\mu_X-\mu_Y$ 的置信区间，其估计公式为：

$$(\bar{x}-\bar{y})\pm t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}}$$

或

$$(\bar{x}-\bar{y})\pm t_{\alpha/2}(n+m-2)\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}$$

例 8-6：某企业有两台生产同一种产品的机器设备，为分析两台机器所生产的产品在重量上的差异，从两台机器所生产的产品中分别随机抽取 11 个和 21 个产品作为样本，并对产品重量进行了测量，根据测量结果计算得到：$\bar{x}=610$ 克，$\bar{y}=595$ 克，$s_x^2=18$，$s_y^2=20$。已知两台机器生产均处正常状态，其产品重量均近似服从正态分布，且方差相同，试构造 $\mu_X-\mu_Y$ 的 95%的置信区间。

解：根据总体方差未知但相等的假设，可以首先计算出方差 σ^2 的估计值 s_p^2：

$$s_p^2=\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}=\frac{(11-1)\times 18+(21-1)\times 20}{11+21-2}=19$$

$\mu_X-\mu_Y$ 的置信区间为：

$$(\bar{x}-\bar{y})\pm t_{\alpha/2}(n+m-2)s_p\sqrt{\frac{1}{n}+\frac{1}{m}}=(610-595)\times 2.042\times\sqrt{19}\times\sqrt{\frac{1}{11}+\frac{1}{21}}=15\pm 3.313$$

根据计算得到，两台机器设备生产的产品的重量差的 95%的置信区间为 11.68～18.31 克之间。

（三）两个总体服从正态分布，总体方差未知且不等

两个总体服从正态分布，总体方差未知且不等时，统计量 $t'=\dfrac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{s_x^2}{n}+\dfrac{s_x^2}{m}}}$ 不再服从自由度为 $n+m-2$ 的 t 分布。解决这一问题的办法是采用修正的方法调整自由度，自由度调整后的公式是：

$$\mathrm{d}f=\frac{\left(\dfrac{s_x^2}{n}+\dfrac{s_y^2}{m}\right)^2}{\dfrac{(s_x^2/n)^2}{n}+\dfrac{(s_y^2/m)^2}{m}}$$

这时统计量 $t'=\dfrac{(\bar{x}-\bar{y})-(\mu_X-\mu_Y)}{\sqrt{\dfrac{s_x^2}{n}+\dfrac{s_x^2}{m}}}$ 近似服从于自由度为 $\mathrm{d}f$ 的 t 分布，这样 $\mu_x-\mu_y$ 的置信区间为：

$$(\bar{x}-\bar{y})\pm t_{\alpha/2}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$$

上例中，若两个总体的方差不等，构造 $\mu_X-\mu_Y$ 的置信区间方法如下。

首先计算自由度：

$$\mathrm{d}f=\frac{\left(\dfrac{s_x^2}{n}+\dfrac{s_y^2}{m}\right)^2}{\dfrac{(s_x^2/n)^2}{n}+\dfrac{(s_y^2/m)^2}{m}}=\frac{\left(\dfrac{18}{11}+\dfrac{20}{21}\right)^2}{\dfrac{(18/11)^2}{11}+\dfrac{(20/21)^2}{21}}\approx 23$$

查 t 分布表，当自由度是 23，置信度是 95%时，t=2.07。代入区间估计的公式得：

$$(\bar{x}-\bar{y})\pm t_{\alpha/2}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}=(610-595)\pm 2.07\times\sqrt{\frac{18}{11}+\frac{20}{21}}=15\pm 3.33$$

即两台机器设备生产的产品的重量差的 95%的置信区间为(11.67, 18.33)。

（四）两个总体不服从正态分布且总体方差未知

对于不服从正态分布的两个总体，往往依据中心极限定理随机抽取大样本，如果两个总体方差未知，就用样本方差 s_x^2 和 s_y^2 分别作为总体方差 σ_X^2 和 σ_Y^2 的估计值，当两个样本容量足够大时，$\mu_X-\mu_Y$ 的置信区间估计公式是：

$$(\bar{x}-\bar{y})\pm Z_{\frac{\alpha}{2}}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}$$

二、两个总体比例差的区间估计

为了估计两个总体比例之差 P_1-P_2，我们可以从每一个总体中各抽取一个随机样本，计算两个样本的比例之差 p_1-p_2，并根据 p_1-p_2 的分布进行置信区间估计。

当两个样本容量 n_1 和 n_2 都很大，且总体比例不太接近于 0 或 1 时，两个独立样本的抽样分布近似地服从正态分布，其期望值为 P_1-P_2，标准差为：

$$\sigma_{p_1-p_2}=\sqrt{\frac{P_1(1-P_1)}{n_1}+\frac{P_2(1-P_2)}{n_2}}$$

即：

$$p_1-p_2 \sim N\left(P_1-P_2,\frac{P_1(1-P_1)}{n_1}+\frac{P_1(1-P_1)}{n_1}\right)$$

由于总体比例为未知参数，所以可以用样本标准差 $s_{p_1-p_2}$ 作为总体标准差的估计量，计算公式是：

$$s_{p_1-p_2}=\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

于是，两总体比例之差的置信区间估计公式为：

$$(p_1-p_2)\pm Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

例 8-7：为了分析新工人和老工人在生产产品质量方面的差异，分别在新职工和老职工生产的产品中随机抽取了 200 件产品和 220 件产品，经检测得到新职工产品的非优等品率为 15%，而老职工产品的非优等品率为 3%，要求在 95%的置信度下估计新老职工产品质量差异的置信区间。

解：根据已知条件 $n_1=200$，$n_2=220$，$p_1=15\%$，$p_2=3\%$，置信度为 95%时，$Z_{\alpha/2}=1.96$，两个总体比例之差的置信区间为：

$$\begin{aligned}&(p_1-p_2)\pm Z_{\frac{\alpha}{2}}\sqrt{\frac{p_1(1-p_{1)}}{n_1}+\frac{p_2(1-p_2)}{n_2}}\\&=(0.15-0.03)\pm 1.96\times\sqrt{\frac{0.15\times 0.85}{200}+\frac{0.03\times 0.97}{220}}\\&=0.12\pm 0.053\end{aligned}$$

即新老职工产品非优等品率之差的 95%的置信区间是(6.7%，17.3%)，说明老职工生产产品的质量明显好于新职工。

三、两个总体方差比的区间估计

在统计实践中，我们通常用方差反映数据内部差异的大小，在上一节中介绍了一个总体方差的置信区间估计。但是实际应用时我们除了要估计一个总体的方差外，还经常会遇到比较两个总体方差的问题。比如，希望比较两个地区内部居民收入差距的大小，就可以用两个地区的收入方差比。

建立两个总体方差比的置信区间，同样需要计算从两个总体中随机抽取的两个样本的方

差比，并根据样本方差比的抽样分布进行估计。

由于两个样本方差比 $\frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F(n_1-1,n_2-1)$ 服从自由度为 n_1-1、n_2-1 的 F 分布，$F_{1-\alpha/2}\leqslant F\leqslant F_{\alpha/2}$。因此可以用 F 分布和样本方差比来构造两个总体方差 σ_1^2/σ_2^2 的置信区间（F 分布图和置信区间如图 8-4 所示）。

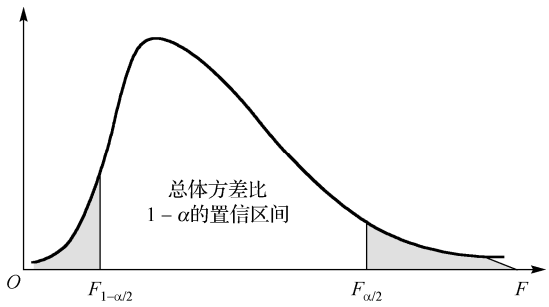


图 8-4　总体方差比的置信区间图

由于 $\frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\sim F(n_1-1,n_2-1)$，于是有 $F_{1-\alpha/2}\leqslant\frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2}\leqslant F_{\alpha/2}$。

据此，可以得到两个总体方差比的置信区间估计公式如下：

$$\frac{s_1^2/s_2^2}{F_{\alpha/2}}\leqslant\frac{\sigma_2^2}{\sigma_1^2}\leqslant\frac{s_1^2/s_2^2}{F_{1-\alpha/2}}$$

第五节　样本容量的确定

前面介绍了一个总体的均值、比例、方差的置信区间估计和两个总体均值之差、比例之差、方差之比的区间估计方法。在估计的公式中，我们注意到要进行总体参数的估计，除了需要计算样本的均值、比例、方差以外，还必须有样本的容量 n。在参数估计的实践之前，其实首先需要解决的就是样本容量 n，即首先解决抽取多大样本的问题。如果抽取的样本过大，就会产生浪费；反之，如果样本太小，又会使估计的误差太大或置信度太低。

一、影响样本容量的主要因素

在参数估计中人们总是希望提高区间估计的置信度，但是在一定样本量和抽样方式下，要想提高置信度就会扩大置信区间，而过宽的置信区间对实际的估计问题又没有意义。例如，如果我们给出居民对某项政策的支持率在 35%～75%之间、地区居民平均收入在 2000～6000 元之间的估计结果，显然没有太大的意义。反过来，如果要缩小置信区间，就会降低置信度，太低的置信度，也同样是没有意义的。当然，如果既要缩小置信区间又提高置信度，就要增加样本容量，因为增加样本容量可以使样本估计量的方差减小，从而使更多的样本集中在总体真值周围。但是样本容量的增加又会带来工作量的增加、费用的增多、时间的延长、非抽样误差的加大等问题，因此样本容量也并不是越大越好。确定合理、适度的样本容量是抽样推断的一个重要问题。从理论上，样本容量的大小主要取决于以下几个因素。

(1)对估计精度的要求，即希望得到的估计值与总体真值之间的离差在什么样的范围以内，或想构造多宽的置信区间。估计区间越大，精度越低，则其他条件相同时，样本容量就越小；估计区间越小，精度越高，则样本容量就越大。

(2)对置信度的要求，即对于规定的置信区间来说想要多大的置信度。置信度与样本容量成正比，当其他条件相同时，置信度越高，样本容量也要越大；而置信度越低，对样本容量的要求也就越低。

(3)总体方差的大小。总体方差是说明总体内部各单位之间差异的大小，其与样本容量成正比关系，总体的方差越大，要求的样本容量越大；反之，如果总体方差越小，则对样本容量的要求也就越小。

此外，实际工作中调查经费和调查时间的限制也会直接影响到样本容量的大小。

二、估计总体均值时样本容量的确定

(一)估计一个总体均值时样本容量的确定

总体均值的置信区间由样本均值和估计误差两部分组成。在重复抽样或无限总体抽样的条件下，估计误差为 $\Delta = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$，$Z_{\alpha/2}$ 的值、总体方差 σ^2 和样本量 n 共同决定了估计误差的大小，一旦确定了置信度 $1-\alpha$，$Z_{\alpha/2}$ 的值就确定了。

对于给定的 $Z_{\alpha/2}$ 的值和总体标准差 σ，可以确定任一允许的估计误差所需要的样本容量，其计算公式是：

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{\Delta^2}$$

如果总体方差未知，可以用以前的经验数据或类似的样本方差来代替，也可以用试调查的办法，选择一个初始样本，以该样本的方差作为总体方差的估计值。

例 8-8：某机构想要估计大学本科毕业生毕业后的起薪水平，打算进行一次随机抽样调查。根据过去经验得知起薪水平的标准差大约为 600 元，要求在 95%的置信度下估计大学本科毕业生的起薪水平，且估计误差不超过 200 元，这样需要样本容量的规模是多大？

解：根据已知条件，$\sigma = 600$元，$\Delta = 200$元，$Z_{\alpha/2} = 1.96$。

代入样本容量的计算公式：

$$n = \frac{\left(Z_{\alpha/2}\right)^2 \sigma^2}{\Delta^2} = \frac{1.96^2 \times 600^2}{200^2} \approx 35\ (\text{人})$$

即应至少随机抽取 35 人进行调查。

上面样本容量的确定是在重复抽样的条件下。当我们采用不重复抽样时，

$$\Delta = Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

从而得到样本容量为：

$$n=\frac{N{Z_{\alpha/2}}^2\sigma^2}{(N-1)\Delta^2+{Z_{\alpha/2}}^2\sigma^2}$$

上例中，若已知某校某年的毕业生人数是 2000 人，采用不重复抽样，则应抽取的人数为：

$$n=\frac{NZ_{\alpha/2}^2\sigma^2}{(N-1)\Delta^2+Z_{\alpha/2}^2\sigma^2}=\frac{2000\times1.96^2\times600^2}{(2000-1)\times200^2+1.96^2\times600^2}=34\text{（人）}$$

即采用不重复抽样时，应至少随机抽取 34 人进行调查。

（二）估计两个总体均值之差时样本容量的确定

对于给定的估计误差 Δ 和置信度水平 $1-\alpha$，估计两个总体均值之差所需要的样本容量为：

$$n_1=n_2=\frac{\left(Z_{\alpha/2}\right)^2\left(\sigma_1^2+\sigma_2^2\right)}{\Delta^2}$$

式中，n_1,n_2 为来自两个总体的样本容量；σ_1^2,σ_2^2 分别为两个总体的方差。

例 8-9：某校为了解男生和女生日常生活费支出水平的差异，打算采取抽样调查的方式进行调查，并对男、女生日常生活费支出之差进行置信区间估计。已知男生和女生生活费支出的方差分别是 8100 和 6400，如果要求估计的置信度为 95%，估计误差不超过 40 元，试确定样本容量。

解：根据已知条件得：$\sigma_{男}^2=8100,\ \sigma_{女}^2=6400,\ \Delta=40$，置信度是 95% 时，$Z_{\alpha/2}=1.96$，则样本容量为：

$$n_1=n_2=\frac{1.96^2\times(8100+6400)}{40^2}=35$$

三、估计总体比例时样本容量的确定

（一）估计一个总体比例时样本容量的确定

在重复抽样或无限总体抽样条件下，总体比例置信区间的估计误差为 $\Delta=Z_{\alpha/2}\sqrt{\frac{P(1-P)}{n}}$。可以看到，$Z_{\alpha/2}$、总体比例 P 和样本量共同决定了估计误差的大小。由于总体比例是固定的，因此估计误差由样本量来确定，样本量越大，估计误差越小，估计的精度越高。从估计误差的角度，我们可以推导出估计总体比例时所需要的样本容量，其计算公式为：

$$n=\frac{\left(Z_{\alpha/2}\right)^2P(1-P)}{\Delta^2}$$

确定样本容量时，如果总体比例 P 未知，可以用类似的样本比例来代替，也可以用试调查的结果，即选择一个初始样本，以该样本的比例作为 P 的估计值。当上面方法无效时，也可以取使 $P(1-P)$ 达到最大值的 P，即 $P=0.5$。

例 8-10：根据以往的经验，某企业生产的产品合格率约为 95%，现要对最近生产的一批产品的合格率进行估计，要求确定当估计误差不超过 5%时的最低样本容量。

解：根据已知条件得：$P=0.95,\ \Delta=0.05,\ Z_{\alpha/2}=1.96$。

将其代入计算公式得：$n=\dfrac{1.96^2\times0.95\times(1-0.95)}{0.05^2}=73$（个）。

若该企业生产的产品合格率未知，也无相关的资料，这时应取 P=0.5，随机抽取的产品数为：

$$n=\frac{1.96^2\times0.5\times(1-0.5)}{0.05^2}=385\text{（个）}$$

（二）估计两个总体比例之差时样本容量的确定

对于给定的估计误差 Δ 和置信度水平为 $1-\alpha$ 时，估计两个总体比例之差时所需的样本容量的计算公式为：

$$n_1=n_2=\frac{\left(Z_{\alpha/2}\right)^2\left[P_1(1-P_1)+P_2(1-P_2)\right]}{\Delta^2}$$

思考与练习

1．简要说明参数的点估计和区间估计的不同及各自的优缺点。

2．如果你看到一条广告，说某药品的有效率为 80%，其误差为正负 3%，那么这条广告给出了什么信息？你相信这条广告吗？如果你要是消费者的话，你还需要追问什么信息吗？

3．如果在置信度不变的情况下，你要使目前所得到的置信区间的长度减少一半，样本量应增加到目前样本量的多少倍？如果保持置信区间长度不变，样本量增加会使结果有什么变化？

4．如果得到总体均值的一个 95%的置信区间为(3.5，4.3)，是否可以说总体均值以95%的概率落入区间(3.5，4.3)之中？为什么？怎样才是合适的说法？

5．有一个调查员问了某超市的 10 个顾客是否喜欢该超市的服务，结果是有 7 个人说喜欢。于是该调查员根据公式 $\hat{p}\pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ 得到喜欢该商店服务的顾客比例的 95%置信区间为(0.42，0.98)。这样做有什么不妥吗？

6．某企业随机抽取了 55 个工人，对其产量进行了统计，其结果如下：

日产量分组(件)	工人数(人)
50～60	5
60～70	13
70～80	19
80～90	10
90～100	8
合计	55

根据上表样本数据，请以 95%的置信度估计该企业职工平均每人日产量的区间。

7．某市人口普查显示，其人口老龄化(65 岁以上)的比率为 14.7%。若你作为某大学暑期社会实践队的成员对该市人口老龄化问题进行研究，随机调查了 800 名该市市民，发现有 114 人年龄在 65 岁以上。那么你的调查结果是否支持该市老龄化率为 14.7%的看法(置信度95%)？

8．某地区有 20 000 亩耕地种植了小麦，调查人员采用不重复抽样方法抽取了其中的

2000 亩，测得平均亩产量为 500 千克，标准差为 125 千克，要求以 95.45%的置信度估计该地区小麦的平均亩产量。

9．为了了解某地区目前居民收入情况，现随机抽取 25 人调查其月收入，得资料如下：

收入(百元)	30 以下	30～40	40～50	50～60	60 以上
人数(人)	2	3	15	3	2

要求：

(1)若该地区居民收入服从 $X \sim N(\mu, 400)$ 的分布，即总体方差为 400，试以 95%的置信水平估计目前该地区居民月收入的可能范围。

(2)若该地区居民收入服从 $X \sim N(\mu, 400)$ 的分布，置信水平为 95.45%，估计总体平均数 μ 的误差不超过 20 元，在简单重置抽样的情况下，最少需要抽取多少人进行调查？

10．现有 100 000 只灯泡，要了解这批灯泡的耐用时间，就随机从中抽取了 400 只，并做了耐用时间的测试和合格检验，测试结果是 400 只灯泡的平均使用时间为 2000 小时，标准差为 12 小时，其中有 8 只不合格。

要求：(1)试计算灯泡耐用时间和合格率的抽样标准误差；

(2)以 95.45%的置信度估计该批灯泡的平均耐用时间；

(3)以 95.45%的置信度估计该批灯泡的合格率。

第九章　参数的假设检验

假设检验与参数估计一样，是通过随机抽取样本并采集样本数据，用样本统计量对总体参数进行推断，是统计推断的主要内容之一。上章所介绍的参数估计是统计推断的主要内容之一，其是通过样本数据去推断未知的总体参数的取值范围和置信度。本章讲述的假设检验也是统计推断的重要的内容和方法，它首先对总体参数的水平提出假设，也就是提出关于总体参数数值的某种说法，然后用样本数据进行验证，做出是否拒绝预先所做出的假设的决策。假设检验可以作为帮助研究人员和管理人员对一种事先的说法进行判断、决策的一种辅助手段，是进行实证分析的重要方法。当然，由于样本的随机性，这种推断也同样具有一定的风险。

第一节　假设检验的基本问题

一、假设检验的基本概念

(一)假设检验的思想

假设检验的基本思想是依据小概率原理，应用反证法，通过观察样本的出现是否属于小概率事件来判别关于总体假设的真伪。

所谓小概率原理，就是认为属于小概率的随机事件在一次试验中几乎不可能发生。如果真的发生了，那么关于这个事件成立的说法(即假设)就值得怀疑。例如，厂商声称，他们生产的产品合格率很高，可以达到 99.9%，如果厂商讲的是真的，即他们厂的产品合格率的确是 99.9%，根据小概率原理，随机抽取的 100 件产品中出现多个(如 5 个)次品的情况几乎不可能发生。但是如果这种情况确实发生了，那么我们就有理由拒绝厂商的说法，不认为他们的产品合格率能够达到 99.9%，或者说其次品率不止 0.1%。

当然，在假设检验中应用反证法具有概率的性质，它只是认为小概率事件几乎不可能发生，而非绝对不可能发生。虽然发生的概率很小，但也不是完全没有可能性。因此，应用概率反证法得出结论时也是存在一定风险的。如企业认为其产品的合格率为 99.9%，从中只随机抽取 1 件产品恰好是次品，0.1%的几乎不可能发生的小概率事件发生了，我们很自然地会拒绝厂商的说法。但是我们也知道次品的概率虽然很小，但也存在 0.1%的可能性，当我们拒绝厂商的说法时，存在有 0.1%的犯错误的可能性。

(二)原假设与备择假设

所谓假设是对总体特征的某种预先判断或陈述。对总体的参数进行假设检验，首先要对研究的总体提出假设。假设检验中，假设包括原假设和备择假设。其中原假设是待检验的假设，假设检验的目的就是检验原假设的真伪，而备择假设则是原假设被拒绝后的替换假设。

(三)检验统计量

对原假设的判断需要根据样本的信息，对随机抽取的样本数据进行加工并用来判断是否拒绝原假设的统计量称为检验统计量。如要检验总体均值是否等于 μ_0（即原假设

$H_0:\mu=\mu_0$），可以计算样本均值$\bar{x}$，若样本均值$\bar{x}$与假设的总体均值μ_0相差很大(若原假设为真，则出现这种情况的概率很小)，我们就有充分的理由拒绝原假设。但是样本均值$\bar{x}$与假设的总体均值μ_0相差多大才是很大，需要进行测量。根据抽样分布理论，当总体服从正态分布[$X\sim N(\mu_0,\sigma^2)$]或从总体中抽取一个大样本时，样本均值服从正态分布，其期望值为μ_0，方差为$\frac{\sigma^2}{n}$，这时衡量两者之间的差异的大小可以用样本均值的标准差作为标准，看两者之间的差是几倍的抽样标准差，即计算$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$作为检验统计量。

检验统计量的选择要根据所研究的参数及其估计量的分布、抽样方式和总体方差是否已知等多种因素来确定。

(四)显著性水平

假设检验的基本原理是小概率原理。所谓小概率原理是指发生概率很小的随机事件在一次试验中几乎是不可能发生的，如果要在一次试验中发生，则表现为检验统计量落入小概率的范围内(即在拒绝域中)，就要拒绝原假设。拒绝域范围的大小取决于小概率的大小，而多大的概率为小概率呢？这需要根据不同的研究对象和要求确定，有的选择 0.05，有的选择 0.01，这一水平通常用α来表示，亦称为显著性水平。显然，α越小越不容易推翻原假设，而一旦拒绝原假设，原假设为真的可能性就越小。假设检验时通常要先确定显著性水平α。

(五)拒绝域和接受域

假设检验是根据检验统计量的具体计算结果来判断是否拒绝原假设H_0，因此在假设H_0为真的情况下将随机抽样的所有可能结果组成的样本空间划分为两部分：一部分是原假设为真时检验统计量极可能(可能性是$1-\alpha$)落入的范围，称为接受域；另一部分是超过了一定的界限，即样本均值与假设的总体均值之间的差异过大，不大可能(可能性是α)出现的范围，即当原假设为真时只有很小的概率才能出现的范围，当检验统计量落入这一范围时便应拒绝原假设，这一区域称为拒绝域。接受域和拒绝域之间的分割点通常称作临界值。

当拒绝域在两侧时[如图 9-1(a)所示]，称为双侧检验；当拒绝域在一侧时[如图 9-1(b)所示]，称为单侧检验。

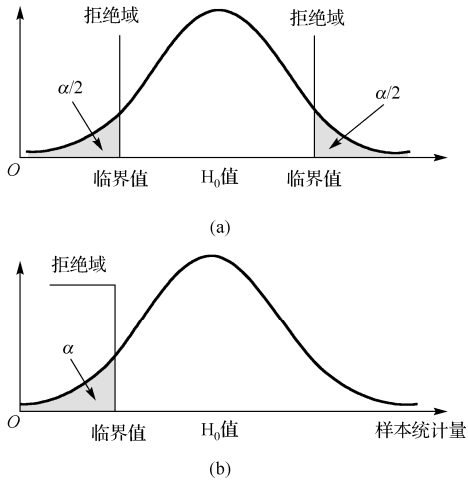


图 9-1 假设检验

二、假设检验的基本步骤

一个完整的假设检验的过程，通常包括以下几个基本步骤。

(一)提出原假设和备择假设

原假设一般用 H_0 表示，备择假设是原假设的对立假设，是在原假设被拒绝后的替换假设，用 H_1 表示。

对于上述两种假设，不能同等看待。原假设作为被检验的假设，应当从已知的原总体(或假设的总体)出发，假设总体没有显著变化，它是不会被轻易否定的，一旦被否定，必有充分的理由。如果原假设没有被拒绝，只能理解为否定原假设的根据还不充分。而不能认定它一定正确。

比如，在正常情况下，零件的平均长度是 2 厘米。现进行了某种技术改造措施，零件的平均长度是否还是 2 厘米呢？这就需要通过随机样本进行检验。这时应首先提出原假设 $H_0:\mu=2$ 厘米，同时提出备择假设 $H_1:\mu\neq 2$ 厘米。在原假设条件下，如果样本的出现属于小概率事件，则有较充分的理由(小概率原理)拒绝原假设，即认为技术改造后，零件的平均长度已不再是 2 厘米。如果样本的出现不属于小概率事件，则不能拒绝原假设。但这只是表明尚无充分的理由否定原假设 $\mu=2$ 厘米，而不认定零件的平均长度一定就是 2 厘米。

以总体均值的假设检验为例，根据实际问题的不同，提出的原假设和备择假设有如下三种类型：

(1) $H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$；

(2) $H_0:\mu\geqslant\mu_0, H_1:\mu<\mu_0$；

(3) $H_0:\mu\leqslant\mu_0, H_1:\mu>\mu_0$。

其中，第一种称为双侧检验，第二种、第三种分别称为左单侧检验和右单侧检验，又统称为单侧检验。

确定原假设和备择假设在假设检验中十分重要，它们直接关系到检验的结果。一般，原假设和备择假设是一个完备的事件组，而且相互对立。这意味着在一项假设检验中，原假设和备择假设必有一个成立，而且只有一个成立。由于原假设所表达的含义总是指参数没有变化或变量之间没有关系，因此等号“=”通常放在原假设，这样一来在双侧检验中，一个总体均值的原假设和备择假设就是 $H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$；在单侧检验中，原假设通常是研究者想收集证据予以推翻的说法，而研究者推翻的假设和研究者支持的假设最终会取决于研究者本人的意向，因而假设的确定带有一定的主观色彩。实际统计检验时，即使是对同一个问题，由于立场不同，研究目的不同，就可能提出截然不同的假设，但是，只要假设的建立符合研究者的目的就是合理的。

如一种零件生产的尺寸标准是 5 厘米，如果零件的平均直径大于或小于 5 厘米，则表明生产过程不正常，需要对生产设备进行调整。这时，研究者想搜集证据予以支持的假设应该是生产过程不正常，即平均直径不是 5 厘米，因为如果研究者认为生产过程正常就不需要进行检验了，所以建立的原假设和备择假设应为：$H_0:\mu=\mu_0, H_1:\mu\neq\mu_0$。

再如，食品厂生产的某种产品的包装上称其净重量为 500 克，地方消协从消费者的利益出发，从产品中随机抽取一批产品来验证其包装说明是否属实。此检验的目的是要保障消费

者的利益，倾向于证明该批产品的净重量不足 500 克，因而建立的原假设和备择假设应为：$H_0:\mu \geq \mu_0, H_1:\mu < \mu_0$。

当然，如果从企业的立场讲，更倾向于证明该批产品净重量大于或等于 500 克，因而建立的原假设应为：$H_0:\mu \leqslant \mu_0, H_1:\mu > \mu_0$。

（二）确定检验统计量

在建立好假设以后是否拒绝 H_0，是根据检验统计量的具体结果是否落入拒绝域而定的。这就要确定什么是检验统计量及该统计量服从什么分布。检验统计量及其分布（包括数学期望和方差）是由许多因素决定的。如检验的是什么参数、总体的分布是否已知和其可能的分布形式，总体的方差是否已知，若检验的参数是两个总体均值之差则还需知道两个总体的方差是否相等等。不同的情况要采用不同的统计量，如 Z 统计量、t 统计量、χ^2 统计量及 F 统计量等。

（三）确定显著性水平和相应的拒绝域

确定显著性水平以后，拒绝域也随之而定。显著性水平的大小应根据研究问题所需的精度而定。如果要求结论比较精确，显著性水平 α 应该小一些；反之，如果要求不太精确，α 可稍大一些，可取 0.05 或 0.1。

（四）做出决策

在规定了显著性水平后，就可以根据原假设与备择假设的设置情况找出拒绝域与接受域的临界点。例如，统计量若服从正态分布，标准化后的统计量为 Z。对 α=0.05 的双侧检验，其临界值为 ±1.96，当统计量 $|Z|>1.96$ 时，就拒绝原假设 H_0，否则没有充分的理由拒绝原假设。

但是，应当注意的是，显著性水平的大小有时会影响假设检验的结果。例如，对同一问题，当显著性水平 $\alpha=0.1$ 时拒绝了原假设，当 $\alpha=0.01$ 时就有可能不拒绝原假设。

三、假设检验中的两类错误

假设检验是根据小概率原理来进行的，因此有可能会判断错误。因为在原假设为真的情况下，有些（只是很少）样本统计量的估计值会落入小概率的拒绝域内而按决策规则加以拒绝。另外，在原假设非真的情况下也有可能有一些统计量的估计值落入接受域的范围之内而没有充分的理由拒绝原假设。因此可以把这些情况归结为两类错误：第一类错误是原假设 H_0 为真而被我们拒绝时，犯了弃真的错误，犯此种错误的概率是 α，所以也叫 α 错误或第一类错误；第二类错误是指原假设 H_0 为伪，而检验的结果不予以拒绝，这类错误被称为取伪错误，发生这种错误的概率通常用 β 表示，故也称为 β 错误或第二类错误。β 值的确定比较困难，需要在决策为不拒绝原假设的条件下，通过具体的概率计算才能知道。

在假设检验中，犯上述两类错误的可能性的大小是相互矛盾的。在样本容量一定的条件下，降低犯弃真错误的可能性（α 降低），必然会增大犯取伪错误的可能性（β 增大）；反之，亦然。要想犯两类错误的可能性都小，必须增大样本容量，而实践中，由于各种条件的限制，样本容量的增大是有限的。

在实际工作中，处理上述矛盾要考虑到实际问题的性质及决策错误所能承担的风险。通

常的做法是控制犯弃真错误的概率，即控制显著性水平 α。根据所研究的实际问题的性质，如果犯弃真错误的损失较大，α 的取值可小些；反之，α 取值可大些。这种控制显著性水平的假设检验被称为显著性检验。

四、假设检验结论的解读

如前所述，拒绝或没有充分理由拒绝一个假设都有可能发生错误。相对而言，做出拒绝判断的理由要充分些，因为通过样本数据我们有充分的理由拒绝(是因为小概率事件发生了)，但接受(没有充分理由拒绝)假设的信心就不是很足。这是因为拒绝一个说法只要找到反例即可，即当前样本就可以说明假设不可靠，而不拒绝某种说法只是表明依据现有的抽取出来的样本还不足以判断假设不成立，只是暂时找不到充分的理由拒绝原假设，但这并不等于以后的观察或试验还找不到证据。因而，在对假设检验结果的表述中尽量避免使用“接受”这样的肯定语气。随着显著性水平取值的减小，拒绝假设的理由将变得更加充分。

假设检验总是在一定的显著性水平 α 下进行的，按照统计频率的思想，显著性水平 α 是指当原假设为真时，在 100 次抽样检验中，求出的检验统计量的值平均有 100α 次落在拒绝域内，即给出 100α 次拒绝的判断，而有 $100(1-\alpha)$ 次不能做出拒绝的判断。

第二节　一个总体参数的假设检验

与参数估计相同，总体参数的假设检验，通常包括有总体均值 μ、总体比例 P 和总体方差 σ^2 的检验。

一、关于总体均值 μ 的假设检验

按总体参数检验的步骤，在确定原假设后采用什么检验统计量取决于总体是否服从正态分布、总体方差 σ^2 是否已知及所抽取的样本是大样本还是小样本等。

(一)大样本检验

按中心极限定理，大样本情况下，无论总体是否服从正态分布，样本均值的抽样分布均近似服从正态分布，其抽样标准差为 $\sigma/\sqrt{n}$。将样本均值 $\bar{x}$ 经过标准化后即可得到检验的统计量。可以证明，样本均值经标准化后得到的 $Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ 服从标准正态分布，而关于原假设的拒绝域与接受域的临界值可查正态分布表。当总体方差 σ^2 未知时，可以用样本方差 S^2 代替。

例 9-1：某汽车生产厂商声称其某种型号汽车的排放量指标的平均水平低于 20 个单位。在抽查了 30 台汽车后之后，得到下面的指标：18.0、20.7、18.9、21.9、20.7、21.4、18.3、22.8、24.2、24.4 …… 22.1，经计算得到该样本均值为 21.13，标准差为 2.1， 现要求在 0.01 的显著性水平下，确认能否由此认为该指标均值超过 20?

解：首先确定原假设和备择假设：

$$H_0:\mu\leqslant 20 \quad H_1:\mu>20$$

由于抽取的是大样本，根据中心极限定理，样本均值近似服从正态分布，总体方差未知

可用样本方差代替，故可以计算出 Z 检验统计量：

$$Z=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{21.13-20}{2.1/\sqrt{30}}=2.94$$

显著性水平 0.01 时，查正态分布表得到临界值 $Z_{0.01}=2.326$。

由于 $Z=2.94>Z_{0.01}=2.326$，落在拒绝域范围内，故可以拒绝原假设，认为该型号汽车的排放量指标高于 20 个单位。

例 9-2：某种袋装食品采用自动生产线包装，每袋的重量标准为 250 克，标准差为 5 克，为检验生产线工作是否正常，质检人员在一批产品中随机抽取了 40 袋进行检验，测得每袋的平均重量为 253 克，取显著性水平 $\alpha=0.05$，检验该生产线包装的袋装食品重量是否正常。

解：此问题关心的是重量是否合格，超过或不足 250 克均不符合要求，因而属于双侧检验，提出的原假设和备择假设为：

$$H_0:\mu=250 \quad H_1:\mu\neq 250$$

由于是大样本，总体方差已知，故可计算 Z 检验统计量：

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{253-250}{5/\sqrt{40}}=3.79$$

查正态分布表得 $\alpha=0.05$ 时 $Z_{0.05/2}=1.96$。

因为 $Z=3.79>Z_{0.05/2}=1.96$，故拒绝原假设，认为袋装食品的重量不符合标准要求。

(二) 小样本检验

在小样本下(通常指 n<30)，进行假设检验时通常首先需要假定总体服从正态分布。至于检验统计量服从什么分布，与总体方差是否已知有关。

当总体方差 σ^2 已知时，由于总体服从正态分布，因而即使在小样本情况下，样本均值标准化后服从标准正态分布，此时可采用服从标准正态分布的 Z 检验统计量进行检验。检验统计量为：$Z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$。

当总体方差 σ^2 未知时，可以用样本方差 s^2 代替总体方差 σ^2。虽然总体服从正态分布，但在小样本情况下，样本均值标准化后不再服从标准正态分布，而是服从自由度为 n–1 的 t 分布，则检验统计量为：$t=\dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}$。

例 9-3：某种零件的长度服从正态分布，质量要求长度为 120 毫米，总体标准差为 5 毫米，为检验生产工作是否正常，质检人员在一批零件中随机抽取了 10 件产品进行了测量，测量结果为：

122　108　120　118　119　124　113　122　120　123

现要在显著性水平 $\alpha=0.05$ 的条件下，检验零件的生产线工作是否正常。

解：根据题意，确定原假设和备择假设：

$$H_0:\mu=120 \quad H_1:\mu\neq 120$$

根据样本数据计算结果得到样本平均数 $\bar{x}=118.9$。

总体服从正态分布，且总体标准差已知为 5 毫米，故可以计算服从标准正态分布的 Z 检验统计量：

$$Z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{118.9-120}{5/\sqrt{10}}=-0.6957$$

$\alpha=0.05$ 时，$Z_{0.05./2}=1.96$。

由于 $|Z|=0.6957<Z_{0.05./2}=1.96$，故不能拒绝原假设，样本提供的证据还不足以推翻原假设，故认为生产线工作正常。

例 9-4：上例中，若总体方差未知，在显著性水平 $\alpha=0.05$ 的条件下，检验零件的生产线工作是否正常。

解：根据题意确定原假设和备择假设：

$$H_0:\mu=120 \quad H_1:\mu\neq 120$$

根据样本数据计算结果得到样本平均数 $\bar{x}=118.9$。

总体服从正态分布，且总体方差未知，根据样本数据计算得到样本标准差 $s=4.932$，故可以计算服从 t 分布的 t 检验统计量：

$$t=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{118.9-120}{4.932/\sqrt{10}}=-0.7053$$

$\alpha=0.05$ 时，$t_{0.05./2}(n-1)=2.262$。

由于 $|Z|=0.7053<t_{0.05./2}=2.262$，故不拒绝原假设，样本提供的证据还不足以推翻原假设，认为生产线工作属于正常状态。

（三）关于总体均值 μ 的假设检验的 SPSS 实现

使用 SPSS 软件，单击【Analyze】下面【Compare Means】中的【One-Sample T Test 项】，然后将要检验的变量选进【Test Variable】中，在【Test Value】中输入原假设中总体均值的 μ_0 的值，再单击【OK】按钮即可。

以下面数据为例：某种零件的长度服从正态分布，按标准要求长度为 120 毫米，为检验生产工作是否正常，质检人员在一批零件中随机抽取 10 件产品进行了测量，测量结果为：

	长度
1	122.00
2	108.00
3	120.00
4	118.00
5	119.00
6	124.00
7	113.00
8	122.00
9	120.00
10	123.00

若总体方差未知，在显著性水平 $\alpha=0.05$ 的条件下，检验零件的生产线工作是否正常。

利用 SPSS 软件进行计算，选择相应的选项（如图 9-2 所示）后出现图 9-3。

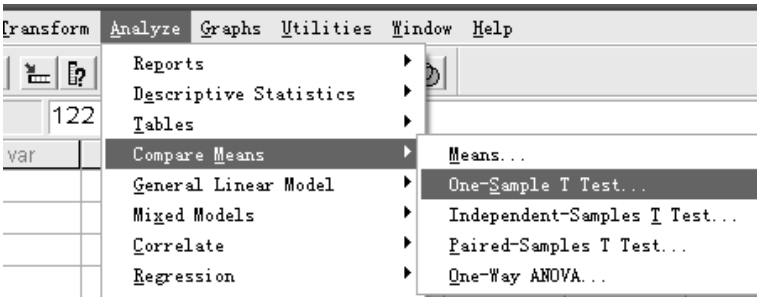


图 9-2 SPSS 操作截图(1)

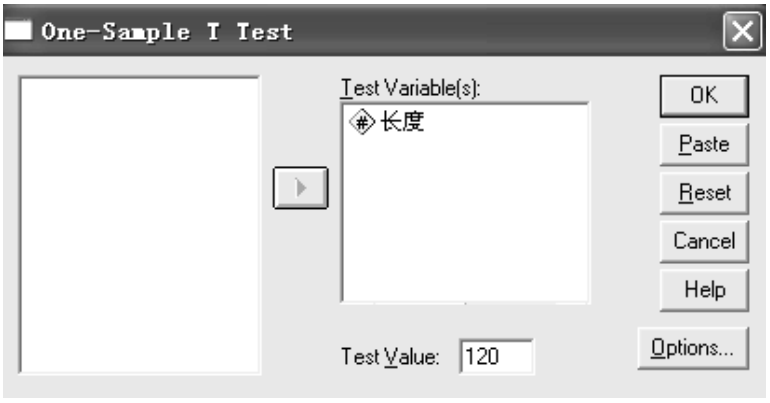


图 9-3 SPSS 操作截图(2)

然后将变量“长度”选进【Test Variable】中，在【Test Value】中输入原假设中总体均值的 μ_0 值，如输入 120，再单击【OK】按钮得到输出结果并解释如下。

(1)关于样本的统计数据。

输出结果如图 9-4 所示，样本容量 n=10，样本均值 118.9，样本标准差 4.93176，样本均值标准差 1.55956。

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
长度	10	118.9000	4.93176	1.55956

图 9-4 输出结果(1)

(2)检验统计量：t=-0.705，自由度 n–1=9。输出结果如图 9-5 所示。

One-Sample Test

	Test Value = 120					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
长度	-.705	9	.498	-1.10000	-4.6280	2.4280

图 9-5 输出结果(2)

(3)检验结果。

检验结论的得出可以有两种方式。

第一是查 t 统计分布表得到一定显著性水平下的临界值，然后根据检验统计量的值 –0.705 与临界值比较，然后得出是否拒绝原假设的结论。

第二是用 P 值决策。从统计上讲，如果原假设 H_0 是正确的，那么我们从该总体中抽取到我们所看到的样本的概率为 P，也称为 P 值，也称为观察到的实际显著性水平(Sig)。当 P(或 Sig)值比较大时，说明大概率事件发生，我们没有理由拒绝原假设。但是如果 P(或 Sig)很小时，说明在原假设条件下的小概率事件发生了，我们对原假设的说法需要质疑。如当 P 值小于所确定的显著性水平 α 时，说明实际观测数据出现的概率比我们想象中的不可能发生的小概率还要小，那就要说明有充分的理由拒绝原假设。上面计算结果(见图 9-5)显示出双侧检验的 Sig 值(即 P 值)为 0.498，没有充分的理由拒绝原假设。P 值与原假设的对或错的概率无关，它是关于数据的概率，也就是说 P 值反映的是来自某个总体的许多样本中某一样本出现的频率高低，它是指当原假设正确时，得到目前这个样本数据的概率。当我们所抽取到的这个样本的概率很小，几乎不太可能发生时，我们就有充分的理由认为原假设是有问题的，从而得出拒绝原假设的结论。

利用 P 值决策的规则很简单，如果 $P<\alpha$，拒绝 H_0；如果 $P>\alpha$，不拒绝 H_0。要注意，SPSS 的这类 t 检验的标准输出都是以双侧检验的 P 值来输出的。因此在解决单侧检验问题时，要把计算机输出的“P 值”减半才能够得到真正的单侧检验的 P 值。

二、关于总体比例 P 的假设检验

对总体比例进行假设检验的思路和程序与总体均值检验类似。在构造总体比例的检验统计量时，通常考虑大样本的情况，利用样本比例 p 与假设的总体比例 P_0 之间的距离的相对大小来衡量，即看样本比例 p 与总体比例 P_0 之差是多少个标准差。由于在大样本情况下样本比例 p 近似地服从正态分布，因而检验统计量样本比例 p 与总体比例 P_0 之间的距离与标准差之比 $Z=\dfrac{p-P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$ 近似地服从标准正态分布。

例 9-5：一项调查结果显示某市老年人口比重为 14.7%，该市老年人口研究会为了检验该项调查是否可靠，随机抽取了 400 名居民，发现其中有老年人 57 人，问调查结果是否支持该市老年人比重占 14.7%的说法？(取 $\alpha=0.05$。)

解：确定原假设和备择假设：

$$H_0: P=14.7\% \quad H_1: P\neq 14.7\%$$

计算样本比例

$$p=\frac{57}{400}=0.1425=14.25\%$$

计算检验统计量：

$$Z=\frac{p-P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}=\frac{0.1425-0.147}{\sqrt{\dfrac{0.147(1-0.147)}{400}}}=-0.254$$

这是一个关于总体比例的双侧检验，当 $\alpha = 0.05$， $Z_{0.05./2} = 1.96$。

因为 $|Z| = 0.0.254 < Z_{0.05/2} = 1.96$，故没有充分的理由拒绝该市老年人口比重为 14.7%的说法。

总体比例的左侧检验和右侧检验与大样本情况下总体均值的检验方法类似。

例 9-6：一个随机抽样的电话调查表明，有 23%的被访问者对某种产品很满意。现在想知道，这是否和企业管理者所期望的总体上至少有 25%的被访问者有显著不足，为此可以进行假设检验。

首先确定原假设和备择假设：

$$H_0 : P \geqslant 0.25 \qquad H_1 : P < 0.25$$

如果 n 为被访问的人数，x 为对产品很满意的人数，那么样本中对产品很满意人数的比例为 $p = x / n = 0.23$。检验统计量则是在零假设下当大样本时近似标准正态的统计量：

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0(1 - P_0)}{n}}} = \frac{0.23 - 0.25}{\sqrt{\frac{0.25(1 - 0.25)}{n}}}$$

这个数值用手算也不费力气。

但是要注意，样本量对于假设检验的结果就十分重要。现在假定 $p = x / n = 0.23$ 不变。

当样本量为 n=1500 时，那么，得到的检验统计量：

$$z = \frac{0.23 - 0.25}{\sqrt{\frac{0.25 \times (1 - 0.25)}{1500}}} = -1.78885$$

P 值为 0.0368。因此，可以认为在确定显著性水平为 0.05 时，说很满意的人数至少 25%是过分了，也就是拒绝零假设。

当样本量为 n=100，那么，得到检验统计量：

$$z = \frac{0.23 - 0.25}{\sqrt{\frac{0.25 \times (1 - 0.25)}{n}}} = -0.046188$$

P 值为 0.3221。因此，在确定显著性水平为 0.05 时，没有足够的理由拒绝 25%的零假设。

利用 SPSS 软件操作的方法如下。

在【Analyze】下拉菜单中，选择【Nonparamitric Tests】中的【Binomial】，然后在【Define Dichotomy】的【Cut point】中输入计算比例时的数据划分的标准(如计算大于 120 的比例，则输入 120 即可)，在【Test Proportion】中输入原假设中的总体比例 P_0，如 0.5，最后单击【OK】按钮即可得到计算结果。

利用软件时，上面的 P 值计算往往在公式中加上用连续变量近似离散变量分布时常用的连续性纠正，因此得出的结果和用上面公式直接手算的稍有不同。

三、关于总体方差 σ^2 的假设检验

方差是测度数据差异程度的量，在统计实践中，仅仅保证观测到的数据的均值在某一正常的水平上并不意味着总体就是正常的。比较均值相同方差不同的产品质量时，方差小的质

量就稳定些，而方差大的产品的质量相对稳定性就差些。与总体方差的区间估计类似，对一个总体方差进行假设检验时，其检验统计量也是使用服从 χ^2 分布的统计量。

进行总体方差 σ^2 的假设检验，其检验的步骤及原理与均值的假设检验基本相同。但是检验时，不论样本量的规模大小，均要求总体服从正态分布，其中检验的统计量为：

$$\chi^2=\frac{(n-1)s^2}{\sigma_0^2}$$

式中，s^2 是样本方差；σ_0^2 为原假设中对总体方差的假设水平。

在给定显著性水平 α 时，双侧检验的拒绝域如图 9-6 所示，单侧检验的拒绝域在分布一侧的尾部。

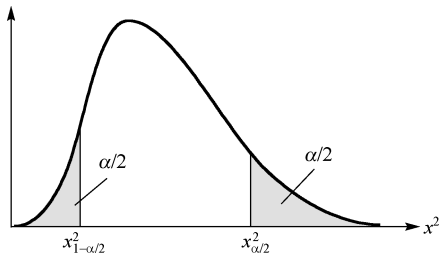


图 9-6　显著性水平 α 时的双侧检验

例 9-7：根据长期正常生产的资料得知，企业所生产的产品的某质量参数服从正态分布，其方差为 0.0025。现从某日生产的产品中随机抽取了 20 个产品，经检验测得样本方差为 0.0042。要求在抽测的基础上，判断该日生产的产品质量参数的波动与平日有无显著差异（$\alpha=0.05$）。

解：首先确定原假设和备择假设：

$$H_0:\sigma^2=0.0025 \qquad H_1:\sigma^2\neq 0.0025$$

根据已知资料 $n=20$，$\alpha=0.05$，自由度 $20-1=19$ 时，$\chi^2{}_{0.025}(20-1)=35.8523$。

计算检验统计量：

$$\chi^2=\frac{(20-1)\times 0.0042}{0.0025}=31.92$$

$$\chi^2=31.92<\chi^2_{0.025}(20-1)=35.8523$$

所以不能拒绝原假设，表明该日产品质量参数的波动与平时相比没有显著差异。

第三节　两个总体参数的假设检验

进行两个总体参数的假设检验主要包括有：两个总体均值之差 $\mu_1-\mu_2$ 的检验、两个总体比例之差 P_1-P_2 的检验和两个总体方差比 σ_1^2/σ_2^2 的检验。两个总体参数的假设检验基本原理与一个总体参数的检验相似，最大的不同在于检验统计量，即由分别从两个总体中随机抽取的两个样本的数据来构造检验统计量。

一、两个总体均值之差的检验

$x_1,x_2,\cdots,x_n$ 是来自均值为 μ_x、方差为 σ_x^2 的总体的一个样本，$y_1,y_2,\cdots,y_m$ 是来自均值为 μ_y、方差为 σ_y^2 的总体的一个样本，且两个样本相互独立。$\bar{x}$、$\bar{y}$ 为样本均值，s_x^2、s_y^2 为样本

方差，α 是检验的显著性水平，要求对两个总体的均值之差 $\mu_x-\mu_y$ 进行检验。

（一）独立大样本条件下的总体均值差的假设检验

根据中心极限定理，在大样本情况下，两个样本均值均近似地服从正态分布，这样两个样本的均值之差 $\bar{x}-\bar{y}$ 也近似地服从正态分布，当两样本独立时：

$$\bar{x}-\bar{y}\sim N\left(\mu_x-\mu_y,\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}\right)$$

则：

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sim N(0,1)$$

双侧检验时：

$$H_0:\mu_x-\mu_y=\mu_0,\ H_1:\mu_x-\mu_y\neq\mu_0$$

构造的检验统计量为：

$$Z=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}$$

拒绝域为：

$$\left|\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\right|>Z_{\alpha/2}$$

单侧左侧检验时：

$$H_0:\mu_x-\mu_y\geqslant\mu_0,\ H_1:\mu_x-\mu_y<\mu_0$$

拒绝域为：

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}<-Z_{\alpha}$$

单侧右侧检验时：

$$H_0:\mu_x-\mu_y\leqslant\mu_0,\ H_1:\mu_x-\mu_y>\mu_0$$

拒绝域为：

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}>Z_{\alpha}$$

如果两个总体的方差 σ_x^2、σ_y^2 未知，可以用样本方差 s_x^2、s_y^2 来代替。

例 9-8：有甲、乙两台设备生产同类型产品，它们生产的产品的某质量参数分别服从方差 $\sigma_{甲}^2=70$，$\sigma_{乙}^2=90$ 的正态分布，现从甲设备生产的产品中随机抽取 35 件，测得该质量参数的平均值 $\bar{x}_{甲}=137$，从乙生产的产品中随机抽取 45 件，测得其平均质量参数 $\bar{x}_{乙}=130$，试问在 0.01 的显著性水平下，能否认为这两台设备生产的产品的该质量参数无显著性差异？

解：根据已知条件，$\sigma_{甲}^2=70$，$\sigma_{乙}^2=90$，$\bar{x}_{甲}=137$，$\bar{x}_{乙}=130$，$n_{甲}=35$，$n_{乙}=45$。

根据问题要求提出原假设和备择假设：$H_0:\mu_x-\mu_y=0$，$H_1:\mu_x-\mu_y\neq 0$。

构造的检验统计量为：

$$Z=\frac{(\bar{x}_{甲}-\bar{x}_{乙})-(\mu_{甲}-\mu_{乙})}{\sqrt{\frac{\sigma_{甲}^2}{n_{甲}}+\frac{\sigma_{乙}^2}{n_{乙}}}}=\frac{(137-130)-0}{\sqrt{\frac{70}{35}+\frac{90}{45}}}=3.5$$

当α=0.01 时，查正态分布表得到临界值为：$z_{\alpha/2}=Z_{0.005}=2.58$。

因为 3.5 > 2.58，说明检验统计量的值落在了拒绝域范围内，故拒绝原假设H_0，说明这两台设备生产的产品的质量参数存在显著差异。

(二)独立小样本条件下的两个正态总体均值差的假设检验

(1)两个总体方差σ_x^2、σ_y^2已知时，无论样本量的大小，从两个正态分布的总体中独立地随机抽取的两个样本均值都服从正态分布，其差值的抽样分布也服从正态分布，这时，用于假设检验的检验统计量为：

$$Z=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}\sim N(0,1)$$

(2)两个总体的方差未知但相等时，即$\sigma_x^2=\sigma_y^2$，则需要用来自两个总体的随机样本的样本方差估计总体方差s_p^2，其估计公式是：

$$s_p^2=\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}$$

这样，检验统计量为：

$$t=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}\sim t(n+m-2)$$

(3)两个总体的方差未知且不相等时，可以用样本方差s_x^2、s_y^2作为总体方差σ_x^2、σ_y^2的估计量，这时检验统计量服从自由度为 v 的 t 分布为：

$$t=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}\sim t(v)$$

其中，自由度的计算公式是：

$$v=\frac{\left(\frac{s_x^2}{n}+\frac{s_y^2}{m}\right)^2}{\frac{\left(\frac{s_x^2}{n}\right)^2}{n-1}+\frac{\left(\frac{s_y^2}{m}\right)^2}{m-1}}$$

自由度取其计算结果四舍五入后的整数。

其拒绝域需要根据检验统计量的分布和假设检验属于单侧检验还是属于双侧检验来确定，确定方法不再赘述。

例 9-9：工厂管理人员对组装新产品的两种方法所需要的时间进行了测试，想了解一下组装的顺序是否影响组装所需的时间，为此，在采用两种组装方法的工人中各随机抽取了12 个工人，测试结果(所用时间：分钟)如下：

甲方法	31	34	29	26	32	35	38	34	30	29	32	31
乙方法	26	24	28	29	30	29	31	26	29	32	28	32

假设两种组装方法所用的时间皆服从正态分布，且方差相同，问在 0.05 的显著性水平下，两种组装方法所需的时间在总体上有无显著差异？

解：根据背景资料可知，两个总体服从正态分布，总体方差未知，利用样本数据计算得到：

$$\bar{x}=31.75,\ \bar{y}=28.67,\ (n-1)s_x^2=112.25,\ (m-1)s_y^2=66.67$$

根据题目要求提出原假设和备择假设：

$$H_0:\mu_x-\mu_y=0,\ H_1:\mu_x-\mu_y\neq 0$$

由于总体服从正态分布，总体方差未知但相等且是小样本，则构造检验统计量为：

$$t=\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}}\cdot\sqrt{\frac{1}{n}+\frac{1}{m}}}=\frac{(31.75\text{-}28.67)-0}{\sqrt{\frac{112.25+66.67}{12+12\text{-}2}}\cdot\sqrt{\frac{1}{12}+\frac{1}{12}}}=2.646$$

当 $\alpha=0.05$ 时，查 t 分布表得 $t_{0.05/2}(22)=2.0789$。

因为 $t=2.646>t_{0.05/2}=2.0789$，故拒绝原假设，认为两种组装方法所需要的时间在总体上有显著的差异。

(三)配对样本数据的假设检验

配对样本数据差值的假设检验，是首先计算配对样本数据的差值 d，然后就差值 d 按单一总体的均值的假设检验方法检验即可。

(四)两个总体均值之差检验的软件实现

1．Excel

在数据分析中根据实际问题选择具体的情况，若两总体方差相等，则选择“t 检验：双样本等方差假设”，若总体方差不等或要进行配对样本的检验则选择相应的选项。单击【确定】按钮后输入变量 1 和变量 2 的区域，并确定显著水平 α 值、确定输出区域后，单击【确定】即可。

例如：两变量(n=m=8)数据资料如下，若检验两变量均值是否有显著差异，则采用 Excel 计算，结果如图 9-7 所示。

	A	B
1	5	6
2	4	6
3	7	7
4	3	4
5	5	3
6	8	9
7	5	7
8	6	6

t-检验：双样本等方差假设

	变量 1	变量 2
平均	5.375	6
方差	2.553571429	3.428571429
观测值	8	8
合并方差	2.991071429	
假设平均差	0	
df	14	
t Stat	-0.72276418	
P(T<=t) 单尾	0.240862696	
t 单尾临界	1.761310115	
P(T<=t) 双尾	0.481725392	
t 双尾临界	2.144786681	

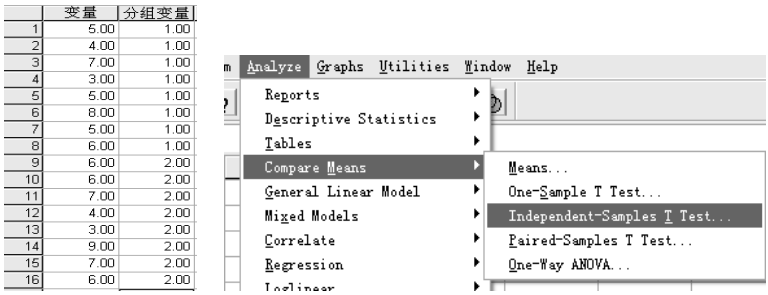
图 9-7　Excel 输出结果

计算结果显示：两变量平均值分别为 5.375 和 6，若两总体方差相同，估计出的共同的方差为 2.99107，自由度为 14，计算 t 检验统计量的值为 –0.72276，由于是检验是否有显著差异，因而属于双侧检验，双侧检验拒绝域的临界值为 2.144787，由于 –0.72276 的绝对值小于 2.144787，故不能拒绝原假设，认为两变量之间不存在显著性差异。若使用 P 值检验，由于双侧检验的 P 值为 0 .4817，大于 $\alpha = 0.05$，故不能拒绝原假设。

2．SPSS

在【Analyze】下拉菜单中选择【Compare Means】，若进行两个独立样本的总体均值检验时，选择【Independent-Samples T Test】，若进行配对样本数据的假设检验则选择【Paired-Samples T Test】，然后选入要进行检验的变量。进行两个独立样本的总体均值检验时，输入用于分组的变量并进行组别的定义；若进行配对样本数据的假设检验，则输入一对变量。单击【OK】按钮即可。

已知数据选项及输出结果如图 9-8 所示。



	变量	分组变量
1	5.00	1.00
2	4.00	1.00
3	7.00	1.00
4	3.00	1.00
5	5.00	1.00
6	8.00	1.00
7	5.00	1.00
8	6.00	1.00
9	6.00	2.00
10	6.00	2.00
11	7.00	2.00
12	4.00	2.00
13	3.00	2.00
14	9.00	2.00
15	7.00	2.00
16	6.00	2.00

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
变量	Equal variances assumed	.003	.956	-.723	14	.482	-.62500	.86474	-2.47967	1.22967
	Equal variances not assumed			-.723	13.707	.482	-.62500	.86474	-2.48340	1.23340

图 9-8　SPSS 输出结果

SPSS 输出结果表明，两组数据之间的方差相等(上面输出表格显示进行等方差检验 P 值(表中为 Sig 值)为 0.956，不能拒绝方差相等的假设)，在方差相等的一行中得到 P 值等于 0.482，远大于 0.05，故不能拒绝原假设，即两变量均值之间不存在显著差异。

二、两个总体比例之差的假设检验

已知 $x_1,x_2,\cdots,x_n$ 为来自二项分布的随机样本，$y_1,y_2,\cdots,y_m$ 为另一个二项分布的随机样本，且相互独立，P_1、P_2 分别为二项分布随机变量取值为 1 的比例，p_1、p_2 为样本比例。np_1、$n(1-p_1)$、mp_2、$m(1-p_2)$ 都大于或等于 10 时，就可以认为是大样本。根据两个样本比例之差的抽样分布，可以得到两个总体比例之差检验的统计量为：

$$Z=\frac{(p_1-p_2)-(P_1-P_2)}{\sigma_{p_1-p_2}}$$

其中，$\sigma_{p_1-p_2}=\sqrt{\frac{P_1(1-P_1)}{n}+\frac{P_2(1-P_2)}{m}}$ 为两个样本比例之差抽样分布的标准差。由于两个总体比例 P_1、P_2 未知，因而需要利用样本比例 p_1、p_2 来估计 $\sigma_{p_1-p_2}$。

(1)若检验两个总体比例之差是否相等，即 $H_0:P_1-P_2=0$，$H_1:P_1-P_2\neq 0$ 时，需要将两个样本合并，并利用合并后的数据计算一个比例代替两个比例，即：

$$p_1=p_2=p=\frac{p_1n+p_2m}{n+m}$$

这样，检验统计量为：

$$Z=\frac{(p_1-p_2)}{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}}$$

(2)若检验两个总体比例之差是否等于某个常数，即：$H_0:P_1-P_2=d$，$H_1:P_1-P_2\neq d$ $(d\neq 0)$ 时，可直接用两个样本比例作为两个总体比例的估计量，这样两个总体比例之差的假设检验统计量为：

$$Z=\frac{(p_1-p_2)-d}{\sqrt{\frac{p_1(1-p_1)}{n}+\frac{p_2(1-p_2)}{m}}}$$

例 9-10：随机抽取了 100 个证券从业人员，其中有 27%的人估计下个交易日股票价格指数将会上升，而同时随机抽取的 200 个普通投资者中有 35%的人认为下个交易日股票价格指数将会上升。调查者认为，证券从业人员认为股票价格指数会上升的比例显著低于普通投资者，问在 0.05 的显著性水平下，能否认为样本提供的证据支持调查者的看法？

解：根据已知条件得：$p_1=0.27$，$p_2=0.35$，确定原假设和备择假设为：$H_0:P_1-P_2\geqslant 0$，$H_1:P_1-P_2<0$。

由于检验的是证券从业人员估计股指会上升的人数所占比例是否显著低于普通投资者，而不是检验两者之间的差值，因而原假设大于等于号的右边是 0，而不是某一不为 0 的常数，因而可以首先计算两个样本的合并比例，即：

$$p=\frac{p_1n+p_2m}{n+m}=\frac{0.27\times100+0.35\times200}{100+200}=0.31$$

这样检验统计量为：

$$Z=\frac{(p_1-p_2)}{\sqrt{p(1-p)\left(\frac{1}{n}+\frac{1}{m}\right)}}=\frac{(0.27-0.35)}{\sqrt{0.31(1-0.31)\left(\frac{1}{100}+\frac{1}{200}\right)}}=-1.72976$$

查正态分布表得 $z_{0.05}=1.645$。

因为 $|Z|=1.72976>Z_{0.05}=1.645$，故拒绝原假设，样本数据提供的证据支持调查者的看法，即证券从业人员认为股票价格指数会上升的比例显著低于普通投资者的比例。

例 9-11：某企业生产产品时进行了技术改造，技术部门声称采用新技术后产品的次品率比原来至少降低了 8 个百分点，为此进行了测试，从新技术生产方法中抽取了 300 个，发现有 33 个次品，从利用原技术生产的产品中也抽取了 300 个，发现有 84 个次品，试在 0.05 的显著性水平下，说明新技术的采纳是否至少使次品率降低了 8%。

解：根据抽样的数据资料计算得到：$n_1=n_2=300,\ p_1=11\%,\ p_2=28\%$。

确定原假设和备择假设：H_0：$P_2-P_1\leqslant8\%,\ H_1:P_2-P_1>8\%$。

由于检验的是两比例之差是不为 0 的常数，因而构造的检验统计量为：

$$Z=\frac{(p_2-p_1)-d}{\sqrt{\frac{p_1(1-p_1)}{n}+\frac{p_2(1-p_2)}{m}}}=\frac{(0.28-0.11)-0.08}{\sqrt{\frac{0.11(1-0.11)}{300}+\frac{0.28(1-0.28)}{300}}}=7.912$$

查正态分布表得 $z_{0.05}=1.645$。

由于 $Z=7.912>Z_{0.05}=1.645$，故拒绝原假设，说明采用新技术生产的产品的次品率至少比原技术生产产品的次品率低 8%。

三、两个总体方差之比的假设检验

已知 $X\sim N(\mu_x,\sigma_x^2),\ Y\sim N(\mu_y,\sigma_y^2)$，且相互独立，$S_x^2$、$S_y^2$ 为样本方差，对 $\frac{\sigma_x^2}{\sigma_y^2}$ 进行假设检验。

(1)若 μ_x、μ_y 已知，则可以用 $s_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)^2$ 和 $s_y^2=\frac{1}{m}\sum_{i=1}^{m}(y_i-\mu_y)^2$ 去估计 σ_x^2、σ_y^2。

这时有：

$$\frac{ns_x^2}{\sigma_x^2}\sim\chi^2(n)\qquad\frac{ms_y^2}{\sigma_y^2}\sim\chi^2(m)$$

因此得：

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}\sim F(n,m)$$

对 $\frac{\sigma_x^2}{\sigma_y^2}$ 进行检验，可以用 $\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ 作为检验统计量。

(2) 若 μ_x、μ_y 未知，则需要用 $s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ 和 $s_y^2=\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\bar{y})^2$ 去估计 σ_x^2、σ_y^2。

这时有：

$$\frac{(n-1)s_x^2}{\sigma_x^2}\sim\chi^2(n-1)\qquad\frac{(m-1)s_y^2}{\sigma_y^2}\sim\chi^2(m-1)$$

因此得：

$$\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}\sim F(n-1,m-1)$$

对 $\frac{\sigma_x^2}{\sigma_y^2}$ 进行检验，可以用 $\frac{s_x^2/\sigma_x^2}{s_y^2/\sigma_y^2}$ 作为检验统计量。

在双侧检验时，通常是用较大的样本方差除以较小的样本方差，这样做是为了能保证拒绝总发生在 F 分布的右侧，所以只需将检验统计量的值与右侧的 $\alpha/2$ 进行比较即可做出判断。

例 9-12：两家企业生产的机械零件的平均使用寿命差别不大，价格也很相近，考虑购买哪家企业生产的机械零件主要取决于其质量的稳定性，即方差的大小，如果方差相同，就选择距离近的企业的零件，如果方差不同，就选择方差小的企业的零件。已知从两个企业生产的零件中各随机抽取 25 个得到样本方差分别为 6.37 和 3.19，假设两企业生产的机械零件的使用寿命均服从正态分布，在显著性水平为 0.05 的条件下进行检验，以确定选用哪一家企业的机械零件。

解：根据已知条件得：$s_x^2=6.37$，$s_y^2=3.19$，$n=m=25$。

提出原假设和备择假设：$H_0:\sigma_x^2\geqslant\sigma_y^2$，$H_1:\sigma_x^2<\sigma_y^2$。

检验统计量为：$\frac{s_x^2}{s_y^2}\sim F(24,24)$。

计算得：$\frac{s_x^2}{s_y^2}=\frac{6.37}{3.19}=1.997$。

$\alpha=0.05$ 时，查 F 分布表得：$F_{0.05}(24,24)=1.98$。

故不能拒绝原假设，认为第二家企业生产的机械零件的方差更小一些，因而应选用第二家企业生产的机械零件。

思考与练习

1．“假设检验的目的是试图接受原假设”的说法对吗？举例说明为什么“不能拒绝原假设”并不等于“接受原假设”。

2．说明假设检验的基本原理。

3．举例说明假设检验的两类错误及其关系。

4．在竞选前针对 A、B 两个候选人进行了民意调查，调查结果显示候选人 A 有 50%的

支持率，而候选人 B 有 48%的支持率；那么是不是候选人 A 在整个选民中的支持率一定大于候选人 B 呢？我们还缺乏什么信息？假定这两个样本量分别为 500 和 1200，你的结论是什么？如果两个样本量均为 5000 呢？

5．为了比较两种鞋底的材料，20 名试验者左右脚穿两种不同材料的鞋，然后记录下左右脚的磨损度。这是独立样本问题吗？如果不是，是什么问题？为什么？利用双尾检验 H_0：$\mu_1=\mu_2$，H_1：$\mu_1\neq\mu_2$，看两种材料的耐磨度是否一样。可选显著性水平为 0.05。

6．显著性水平是否是原假设正确的概率？如果不是，如何解释？

7．在做出任何结论时都应该给出你的结论可能犯错误的概率，假设检验中，如果是拒绝原假设，那么可能犯什么错误？犯错误的最大可能性是多大？

8．一种元件，要求其使用寿命不低于 1000 小时。现从中随机抽取 25 件，测得其平均寿命为 950 小时。已知该种元件寿命服从标准差为 100 小时的正态分布，试在显著性水平为 0.01 的要求下判断这批元件是否合格？

9．已知某零件的尺寸服从正态分布，现从某天生产的零件中随机抽取 10 个，测得其长度(毫米)如下：

14.8	15.1	14.6	15.2	14.9
15.0	14.8	15.1	15.3	14.7

要求：

(1)确定该种零件平均长度的置信区间，置信水平$1-\alpha=95\%$。

(2)若要求该种零件的标准长度应为 15 毫米，试在显著性水平$\alpha=0.05$条件下，检验该种零件是否符合标准要求。

10．甲、乙两厂生产同种零件，已知零件长度均服从正态分布，且$\sigma_{甲}^2=400$，$\sigma_{乙}^2=529$。从甲厂生产的零件中随机抽取了 81 件，测得$\bar{x}_{甲}=400$厘米，从乙厂生产的零件中随机抽取了 100 件，测得$\bar{x}_{乙}=420$厘米。根据以上调查结果，能否认为甲、乙两厂生产的零件平均长度相等？

11．用两种方法生产组装产品，为比较两种方法组装效率是否有显著差异，现随机独立抽取两组各 12 人进行测试得到数据(件/天)如下：

原方法	新方法	原方法	新方法
28	27	36	31
30	22	37	26
29	31	38	32
37	33	34	31
32	20	28	33
28	30	30	26

若新旧方法方差相等，在 5%的显著性水平下能否认为新旧方法组装产品的劳动生产率相等？

利用 Excel 统计功能计算如下：

t-检验：双样本等方差假设	变量 1	变量 2
平均	32.25	28.5
方差	15.47727273	18.45454545
观测值	12	12

续表

t-检验：双样本等方差假设	变量 1	变量 2
合并方差	16.96590909	
假设平均差	0	
df	22	
t Stat	2.230069126	
P(*T*<=*t*) 单尾	0.018132009	
t 单尾临界	1.717144335	
P(*T*<=*t*) 双尾	0.036264019	
t 双尾临界	2.073873058	

第三部分　变量之间关系的统计推断

社会经济与自然科学现象之间的相关联系和制约是一个普遍规律，社会经济的发展总是与一定的经济变量的数量变化紧密联系。在经济现象的内部和外部联系中存在着一定的相关性。研究变量之间的关系并利用这种关系，可以有助于我们认识客观事物发展变化的规律，帮助我们进行客观的预测和科学的决策，指导并控制社会经济活动的发展。

变量之间的关系包括定性变量与定性变量之间的关系、定性变量与数值型变量之间的关系和数值型变量与数值型变量之间的关系等。

第十章　方 差 分 析

在第九章中我们介绍了如何通过统计检验来判断两个总体之间的参数是否有显著差异。但在统计分析实践中除了需要分析与比较两个总体之间的参数的差异，还经常会遇到多个总体参数的比较与分析。例如，检验多个总体均值是否相等的问题，对此，我们可以用前面两总体参数是否有显著差异的假设检验的方法，对多个总体的参数进行两两比较分析和统计检验。当然，如果总体个数较多，分别进行两两总体之间的检验就显得比较麻烦，这样就要考虑能不能同时对这些总体的均值进行检验的问题。

上述所说的多个总体，也可理解为多个不同类别，即按某分类变量将研究对象划分为不同的类别，继而检验各类(总体)之间是否有显著差异，此问题也可以理解为分类变量对参数(数值型变量)是否有显著影响。要分析多总体之间的参数是否存在显著差异，或分析分类变量对总体数值型变量是否有显著影响 ，我们可以通过方差分析(Analysis of Variance，ANOVA)进行。

从形式上看，方差分析是比较多个总体的均值是否相等的方法，本质上研究的是分类变量与数值型变量之间的关系。其广泛应用于心理学、生物学、工程和医药试验数据等领域与专业。

第一节　方差分析的基本原理

为了解方差分析的目的、基本思想与原理，我们首先分析一个简单的问题。

一、问题的提出

例 10-1：某食品公司研制出一种新的产品，想了解这种新的食品上市后是否会因为销售位置不同而存在销售差异。为此，该公司在某城市城东、城西、城南、城北四个区域分别选取了经营规模相近的五家超市上进行了试销，并收集了其销售量数据(如表 10-1 所示)。

表 10-1　该饮料在五个地区的销售情况(袋)

超市	城东	城西	城南	城北
1	312	265	279	308
2	283	287	251	296
3	308	251	285	324
4	279	291	242	317
5	296	272	265	328
均值	295.6	273.2	264.4	314.6

从表 10-1 中的数据可以看到，这四个区域的销售量均值都不相同，能否推断得出不同区域的消费者对该食品的偏好不同，即不同区域该食品销售量是否存在显著差异？或者说这种不同到底是否是由区域因素造成的？

在其他条件相同的情况下，上述问题就归结为一个检验问题，即：区域因素对销售量是否有显著影响？

为此，提出检验的假设为：

$H_0:\mu_1=\mu_2=\mu_3=\mu_4$　区域因素对销售量无显著影响

$H_1:\mu_1,\mu_2,\mu_3,\mu_4$ 不全相等　区域因素对销售量有显著影响

μ_i 表示各区域销售的平均数。

要检验总体上这四个区域该产品的销售量均值是否相等。对此，我们可用方差分析的方法来判断。

二、方差分析的原理及应用条件

下面我们以区域位置因素是否影响食品销售水平问题为例介绍方差分析的原理。

从表10-1中的数据我们可以看到，不同区域(城东、城西、城南、城北)及不同超市上的销售量量多多少少都存在一定的差异，导致这种差异的原因是由于区域的原因，还是由于其他的随机因素需要我们分析，如果这种差异是由于区域的不同导致的，我们就会得出区域是影响销售的因素。对此我们需要对这种数据间的差异进行分析。

首先，针对本例，我们可以用表10-1中的样本数据资料测量所有数据的差异程度，即计算出所有区域该食品销售量与销售量均值总的离差平方和 SST，即：

$$\mathrm{SST}=\sum_{i,j}(x_{ij}-\bar{\bar{x}})^2=11592.95$$

式中，$\bar{\bar{x}}$ 是所有数据的平均数；x_{ij} 是各区域在各超市的销售量。

由计算公式可知，总离差平方和反映了在每个区域每一家超市该食品的销售量与总平均数的差异程度。为分析这一差异的形成原因，我们需要对反映数据差异的总离差平方和进行分解，将本例中四个区域的销售数据看作为四个组，将销售量的离差总平方和分解为以下两个部分。

一部分是每一组组内(即同一区域内)各超市的食品销售量与本组的平均销售量的差异程度，称为组内差的平方和：$\sum_j(x_{ij}-\bar{x}_i)^2$，其中 $\bar{x}_i$ 表示第 i 区域的平均销售量。

将四个组的组内差的平方和相加即得到总的组内差的平方和 $\mathrm{SSE}=\sum_i\sum_j(x_i-\bar{x}_i)^2$。这部分差是在相同区域、相同产品、超市的经营规模环境等都近似的情况下的销售量数据之间的差异，其主要是由于样本的随机性而产生的，即随机误差。

另一部分是组与组之间销售量的差异程度，称为组间离差平方和：$\mathrm{SSA}=\sum_j n_i(\bar{x}_j-\bar{\bar{x}})^2$。这一差异来自两个方面：一是由于样本的随机性而产生的差异；二是由区域的不同所造成的，即销售区域的不同对销售量产生了影响。

如果销售区域对销售结果没有影响，那么在组(区域)之间的差异中，就仅仅有随机因素的差异，而没有系统性差异(区域不同产生的差异)，它与组内部的差异就应该接近。反之，如果组间差异(不同区域间的差异)很大，远远大于组内部的差异，那么就认为组间差异受销售区域的影响。因此我们只要将组之间的差异与组内部的差异进行对比，就可以知道是否有系统性差异(销售区域不同产生的差异)。

为了比较组(区域)之间的差异与组(区域)内部的差异是否显著，我们可以计算检验统计

量。由于上面组间总离差平方和与组内总离差平方和不能直接进行对比，我们需要计算其各自的平均差异程度，即组间方差 $\mathrm{MSA}=\dfrac{\sum_j n_i(\bar{x}_j-\bar{\bar{x}})^2}{k-1}$ 和组内方差 $\mathrm{MSE}=\dfrac{\sum_i\sum_j (x_{ij}-\bar{x}_i)^2}{n-k}$，其中，$k$ 为分组的组数。将这两个方差进行对比所形成的统计量，服从自由度为（$k-1,n-k$）的 F 分布，称为 F 统计量，即：

$$F=\frac{\text{组间方差}}{\text{组内方差}}$$

F–分布密度[以自由度为(3，20)和(50，20)为例]曲线如图 10-1 所示。

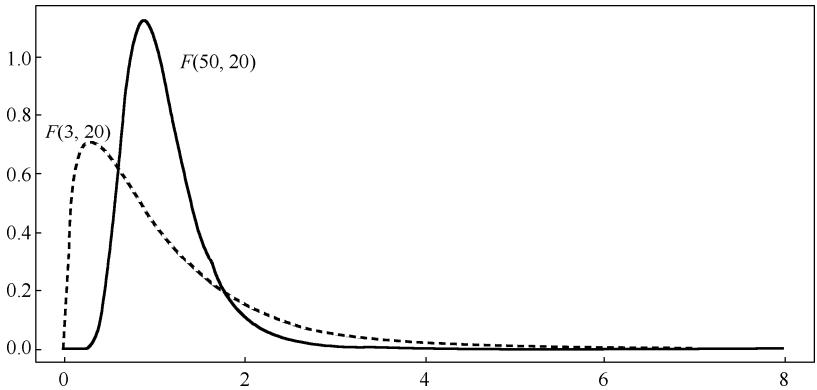


图 10-1　F–分布密度曲线图

由 F 分布图可知当计算出来的 F 统计量数值较大时，其对应的概率就会很小。当出现小概率时，我们就认为组间差异中的区域因素产生的作用是显著的。

从上面的分析中我们看到，利用方差分析进行影响因子对因变量影响作用的显著性检验，构造了 F 检验统计量。在应用时需注意，数据必须满足以下三个假定条件。

(1)所有总体的因变量(如每个区域的销售量)均服从正态分布。对于本例来说，就是每个区域的观测值是来自正态总体的简单随机样本。

(2)所有总体的因变量的方差相等。本例中，各区域观察数据，是从具有同方差的正态总体中抽取的，即各区域销售量的方差相等。

(3)因变量各观察值之间相互独立。本例中，每个被抽中的超市的销售量与其他超市的销售量之间是相互独立的。

三、方差分析中的基本概念

(1)因变量(Dependent Variable)是我们要研究的变量。在上例中，我们要分析的是区域因素对销售量的影响是否显著，其中销售量就是我们所关心的变量，即因变量。

(2)因素或因子(Factor)是研究中所要考虑和分析的影响因素，是能够人为进行控制的因素，也称因子或自变量，本例中的区域因素就是因子。方差分析的因子应该是定性变量。

(3)因子的水平或处理(treatment)。为了研究因子对因变量的影响，需要考察因子的两个或更多个不同的取值情况，因子的这些不同表现或取值称为因子的水平或处理。本例中，区域因子的水平有四个：城东、城西、城南和城北。

(4) 观测值。即每个因子水平下得到的样本数据。

(5) 随机误差。随机误差是指由于抽样的随机性造成的数据之间的差异。例如，在同一区域内，所抽取的样本(超市)的各观测值是不同的，他们之间的差异可以看成是由随机因素的影响造成的。

(6) 系统误差。系统误差是指数据之间的差异是由于因子水平不同而产生的。例如，在不同的区域，其销售量的平均水平不同，主要是由于区域因素造成。

根据因变量、影响因子等的数量与特点，可以对方差分析方法有不同角度的分类。

(1) 按研究变量即因变量的多少，分为一元方差分析和多元方差分析，本章只讨论一元方差分析。

(2) 按影响因素即因子的多少，分为单因素方差分析和多因素方差分析。

(3) 按是否考虑协变量，是否考虑定量变量对因变量的影响，分为不考虑协变量的方差分析和考虑协变量的方差分析。

在一元方差分析有且只有一个因素的情况下，涉及两个变量：一个是分类型自变量；另一个是数值型的因变量。

第二节　一元单因素方差分析

当方差分析中只涉及一个数值型因变量和一个分类型自变量时，称为一元单因素方差分析。它用于检验一个影响因子是否对因变量具有显著的影响作用，即统计上的显著作用，还可以对该因素的若干水平分组中哪些组之间具有显著性差异进行分析。方差分析时，要求因变量数据来自于正态分布总体，如果因变量的分布明显是非正态分布，方差差异大，或数据缺乏独立性，则不宜使用方差分析。

一、数据结构形式

用 A 表示因素，该因素的 K 个水平分别用 $A_1, A_2, \cdots, A_k$ 表示，每个观测值用 x_{ij} 表示，即第 i 个水平(总体)的第 j 个观测值，具体结构表示如表 10-2 所示。

表 10-2　一元单因素方差分析数据结构表

观测值	因素 A			
	A_1	A_2	…	A_k
1	x_{11}	X_{21}		x_{k_1}
2	x_{12}	x_{22}		x_{k_2}
…				
…				
n	x_{1n_1}	x_{2n_2}		x_{nk_k}

二、数据分析与计算步骤

为了检验自变量因素对数值型因变量是否有显著影响，或多个总体均值之间是否存在显著差异，需要做如下的计算、分析和检验。

(一) 提出假设

方差分析的目的是要检验定性因子对因变量是否有显著影响，或多个总体均值之间是否存在显著差异，即检验因子的 K 个水平(总体)的均值是否相等。为此需要提出如下假设：

$H_0: \mu_1 = \mu_2 \cdots = \mu_k$ 定性因子对因变量无显著影响

$H_1: \mu_1, \mu_2 \cdots \mu_k$ 不全相等 定性因子对因变量有显著影响

如果无法拒绝原假设 H_0，则不能认为定性因子变量对因变量有显著影响，也就是说不能认为因子变量与因变量之间有关系；如果拒绝原假设，则意味着定性因子变量对因变量的变化有显著影响。当然，拒绝原假设 H_0，只能表明至少有两个总体的均值不相等，而不能得出所有均值均不相等的结论。

(二)构造并计算检验统计量

根据前述的方差分析原理，我们需要将组间方差 $\text{MSA} = \dfrac{\sum_j n_i(\bar{x}_j - \bar{\bar{x}})^2}{k-1}$ 和组内方差 $\text{MSE} = \dfrac{\sum_j \sum_i (x_i - \bar{x}_i)^2}{n-k}$ 进行对比，对比得到的检验统计量服从自由度为 $(k-1, n-k)$ 的 F 分布(其中 n 为所有观测数据个数，k 为水平或分组的个数)。下面结合上例中数据介绍检验统计量的计算过程。

(1)首先计算各水平(总体)的均值 $\bar{x}_i$。

$$\bar{x}_i = \frac{\sum_j x_{ij}}{n_i} \qquad (i = 1, 2 \cdots k)$$

其中，n_j 为第 j 个水平的样本观测数据个数。

本例中的各水平(总体)的均值 $\bar{x}_i$ 计算结果如表 10-3 所示。

表 10-3 该饮料在五个地区的销售均值(袋)

超市	城东	城西	城南	城北
1	312	265	279	308
2	283	287	251	296
3	308	251	285	324
4	279	291	242	317
5	296	272	265	328
$\bar{x}_j$	295.6	273.2	264.4	314.6

(2)计算全部观测值的总均值 $\bar{\bar{x}}$。

$$\bar{\bar{x}} = \frac{\sum_{i,j} x_{ij}}{n} = 286.95$$

其中，n 为所有样本数据个数。

(3)计算总离差平方和，并进行总离差平方和的分解。

根据方差分析的基本原理，构造检验统计量需要计算总离差平方和、水平间(组间)差的平方和和水平内(组内)差的平方和。

总离差平方和 SST，即全部观测值 x_{ij} 与总均值 $\bar{\bar{x}}$ 值的离差平方和，用以反映全部观测数据的离散状况，也称为总平方和，计算公式是：

$$\mathrm{SST}=\sum_i\sum_j(x_{ij}-\bar{x}_i)^2\text{，本例中 SSA}=11592.95$$

水平间(组间)离差平方和 SSA，即各组平均值与总平均值的离差平方和，反映各总体之间的差异程度，也称为组间平方和。组间平方和既包含组间的差异，也包括样本的随机误差，其计算公式是：

$$\mathrm{SSA}=\sum_i\sum_j(\bar{x}_i-\bar{\bar{x}})^2=\sum_i n_i(\bar{x}_i-\bar{\bar{x}})^2\text{，本例中 SSA}=7684.55$$

水平内(组内)离差平方和 SSE，它是每个水平或组的各样本数据与其组平均误差的平方和，反映了水平内部各观测值的离散状况。组内平方和实际上反映的是随机误差的大小，其计算公式是：

$$\mathrm{SSE}=\sum_i\sum_j(x_{ij}-\bar{x}_i)^2\text{，本例中 SSE}=3908.4$$

上面三个离差平方和之间存在如下等式关系：

$$\sum_i\sum_j(x_{ij}-\bar{\bar{x}})^2=\sum_i n_i(\bar{x}_j-\bar{\bar{x}})^2+\sum_i\sum_j(x_{ij}-\bar{x}_i)^2$$

总离差平方和 = 水平间离差平方和 + 水平内离差平方和

$$\mathrm{SST}=\mathrm{SSA}+\mathrm{SSE}$$

三个离差平方和中，SSE 是对组内的离差即样本的随机误差大小的度量，它反映了除影响因子对因变量的影响外，其他因素对因变量的总影响，因此 SSE 也被称为残差变量或残差效应；SSA 是对样本的随机误差和影响因子导致的系统差异的大小的度量，它不仅包含随机影响因素的作用，也反映了影响因子(区域因素)对因变量(销售量)的影响，也称为影响因子效应；SST 是对全部数据总差异程度的度量，是 SSA 与 SSE 之和。

如果原假设成立，则表明影响因子对因变量没有作用，即水平间的差异与水平内的差异相当，如果水平间与水平内差异显著，则说明影响因子对因变量水平的差异有显著影响。

当然，由于离差平方和的大小与观测数据个数的多少有关，故在比较时需要计算方差。

(4)计算检验统计量。

为了消除观测数据个数的多少对离差平方和的影响，需要将其进行平均，用各平方和除以他们所对应的自由度，得到相应的均方差。

SST 的自由度为 $n-1$;

SSA 的自由度为 $k-1$；

SSE 的自由度为 $n-k$。

因此：

$$\text{水平间均方差 MSA}=\frac{\mathrm{SSA}}{k-1}\text{，本例中 MSA}=2561.517$$

$$\text{水平内均方差 MSE}=\frac{\mathrm{SSE}}{n-k}\text{，本例中 MSE}=244.275$$

将 MSA 与 MSE 进行对比，即可得到检验统计量，当原假设为真时，该统计量服从自由度分别为 $k-1$和$n-k$ 的 F 分布，即：

$$F=\frac{\mathrm{MSA}}{\mathrm{MSE}}\sim F(k-1,n-k)$$

本例中：$F=10.486$。

(5)做出判断。

计算出检验统计量 F 值后，将其与显著性水平 α 下的临界值进行比较，进而作出判断得到结论。根据我们事前确定的显著性水平 α，在 F 分布表中查找自由度分别为 $k-1$ 和 $n-k$ 的 F 分布表中的临界值 $F_\alpha(k-1,n-k)$。

若 $F>F_\alpha$，则拒绝原假设 H_0，即 $H_0:\mu_1=\mu_2\cdots=\mu_k$ 不成立，说明各水平间均值的差异是显著的，或影响因子对因变量的影响是显著的。

若 $F<F_\alpha$，则没有充分的理由拒绝原假设，不能认为各水平间的差异是显著的，或者说不能认为影响因子对因变量的影响是显著的。

除了可以利用 F 分布表中的临界值 $F_\alpha(k-1,n-k)$ 进行比较从而得出检验的结果，我们也可以用实际的概率 P 值与显著性水平 α 比较，若 P 值小于 α，则拒绝原假设，否则不能拒绝原假设。

三、关系强度的测量与多重比较

(一)关系强度的测量

上面方差分析的结果表明，不同区域之间的销售量存在着显著的差异，或说区域因子对销售量有显著的影响。既然区域因子对销售量有显著的影响作用，那我们要进一步分析该影响因子对销售量的影响强度，可以利用总离差平方和及其构成。当总离差平方和一定时，水平间的离差平方和越大，说明影响因子变量对因变量的影响强度越大。

若要度量影响因子变量对因变量的影响强度，我们可以计算水平间的离差平方和占总离差平方和的比重，即：

$$R^2=\frac{\mathrm{SSA}}{\mathrm{SST}}$$

本例中，$R^2=\dfrac{\mathrm{SSA}}{\mathrm{SST}}=\dfrac{7684.55}{11592.95}=66.29\%$，表明区域因素对销售量的影响效应占销售量总差异的 66.29%，也就是说区域因素能对销售量差异的解释比例达到 66.29%。

其平方根 R 就可以用于度量两个变量之间的关系强度。

(二)多重比较分析

前面所进行的方差分析的举例分析中，得到的结果是拒绝原假设，即不同区域的销售量均值之间存在显著性差异，但是我们不知道是哪一区域或哪几个区域的销售量与其他区域的销售量有显著差异。因为方差分析中只要有一个区域的销售量均值与其他任何一个区域不同，就可以拒绝原假设。也就是说拒绝原假设并不表明任何两个水平之间都有差异。因此有时候还需要研究哪些均值之间有差异，哪些均值之间无差异。同时进行两两总体均值之间有无差异的问题称为多重比较问题。

在进行多重比较分析时有方差相等假设条件和方差不等假设条件下的多重比较方法。以下只介绍最常用的方差相等条件下的多重比较方法。

方差相等假设下的多重比较方法常用的是 LSD 法，是 Fisher(费雪)提出的最小显著性检验(Least Significant Difference，LSD)方法，适用于总体方差相等的情况，而且对重复试

验次数相同和不同的情况都适用。该方法运用 t 检验进行任意两个均值之间的配对比较，使用该方法进行检验的具体步骤如下。

(1)提出假设 H_0：$\mu_i = \mu_i$, H_1：$\mu_i \neq \mu_j$。

(2)计算检验统计量：$\overline{x}_i - \overline{x}_j$。

(3)计算检验临界值 LSD，其计算公式是：

$$\mathrm{LSD} = t_{\alpha/2}\sqrt{\mathrm{MSE}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

式中，$t_{\alpha/2}$ 为自由度 $n-k$ 的 t 分布的值；n_i、n_j 分别为因子水平为 i 和 j 的样本的容量，即数据个数。

(4)用检验统计量 $\overline{x}_i - \overline{x}_j$ 与检验临界值 LSD 进行比较，若 $\left|\overline{x}_i - \overline{x}_j\right| > \mathrm{LSD}$，则拒绝原假设，说明 i 和 j 两个因子水平下的均值有显著差异；否则不拒绝原假设。

四、方差分析的软件操作与实现

从上面的计算分析过程可以看到，如果手工计算则需要比较大的计算量，比较麻烦，对此，我们可以利用统计软件来实现操作。以前面区域因子对销售量影响的分析为例，要分析其对销售量的影响是否显著，或不同区域间的销售量是否存在显著差异，现利用 SPSS 软件进行方差分析的相关计算，过程与操作方法如下：

SPSS 选项：【Analyze】—【Compare Mean】—【One-Way ANOVA】。

第一步：将变量“销售量”放入【Dependent List】中；将因子变量“区域”放入【Factor】中；

第二步：单击【Options】按钮进入，选择【Homogemety of Variance Test】(进行方差分析的假定条件：所有总体的研究变量的方差相等的检验)项，选择【Means Plots】(画均值图)项；

第四步：单击【Post Hoc】按钮进入，选择【Equal Variances Assumed】中的“LSD”(进行多重比较检验)。

(一)基本假定的检验

前面已经说过，方差分析中，对试验结果有三个基本假定：(1)正态性；(2)方差一致性(方差相等，如图 10-2 所示)；(3)相互独立性。在实际问题的分析中，如果数据不能满足这三个假定，则方差分析中的统计量不再服从 F 分布，方差分析的结论将不可靠。

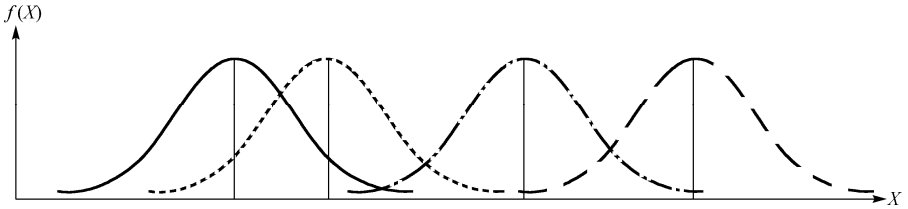


图 10-2　方差相同的分布

若在获取数据的过程中很好地遵循了随机原则，调查结果的相互独立性就不成问题。而正态性检验就不那么简单，当样本量较小时，难以直观判断和检验资料是否来自正态总体，

因而常常需根据过去的经验作出判断。当样本量较大时，无论资料是否来自正态分布总体，数理统计的中心极限定理保证了样本均值的抽样分布仍然服从或渐近服从正态分布。由于正态性与方差一致性往往是伴随的，即不服从正态分布的资料往往方差也不一致，因此，在进行均值比较时对小样本资料进行方差一致性检验就显得尤为重要。方差的一致性常用 Levene 方法检验。

Levene 方法能够对两组或多组资料进行方差一致性检验，不依赖于总体分布的具体形式，适合于任意分布，比其他方法更稳健。Levene 检验在国外已被广泛认同为方差一致性检验的标准方法。

本例中我们需要做方差一致性的检验，其检验的假设为：

$$H_0:\ \sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2$$

H_1：上述方差不全相等

Levene 检验的输出结果如表 10-4 所示。

表 10-4 方差齐性检验

Levene 统计量	第一自由度	第二自由度	*P* 值
.282	3	16	.838

可以看出，Sig(即 *P* 值)＝0.838＞0.05，即在 5%的显著性水平下，没有足够理由拒绝四个总体之间的方差相等的原假设，符合方差分析关于方差相等的前提条件。

(二)方差分析结果

方差分析结果如表 10-5 所示。

表 10-5 方差分析——销售量

	平方和	自由度	均方差	*F* 值	*P* 值
水平间	7684.55	3	25.615	10.486	.000
水平内	3908.40	16	2.443		
Total	11592.95	19			

表中水平间平方和水平内平方和分别为 7684.55 和 3908.40。水平间方差、水平内方差分别为 2561.50 和 244.30。*F* 值为 10.486，Sig(*P* 值)=0.000，是小概率，应拒绝 $\mu_1=\mu_2=\mu_3=\mu_4$ 的原假设。即在 0.05%的显著性水平下认为不同区域的销售量均值之间有显著性差异，从而得出区域因子对销售量有显著性影响的结论。

(三)多重比较分析

本例中的多重检验的计算结果如表 10-6 所示。

表 10-6 多重检验的计算结果

多重比较						
		均值差(*I*–*J*)	标准误	显著性	95% 置信区间	
					下限	上限
东	西	22.400*	9.885	0.038	1.45	43.35
	南	31.200*	9.885	0.006	10.25	52.15
	北	–19.000	9.885	0.073	–39.95	1.95

续表

		均值差(*I*–*J*)	标准误	显著性	95% 置信区间	
					下限	上限
西	东	–22.400*	9.885	0.038	–43.35	–1.45
	南	8.800	9.885	0.387	–12.15	29.75
	北	–41.400*	9.885	0.001	–62.35	–20.45
南	东	–31.200*	9.885	0.006	–52.15	–10.25
	西	–8.800	9.885	0.387	–29.75	12.15
	北	–50.200*	9.885	0.000	–71.15	–29.25
北	东	19.000	9.885	0.073	–1.95	39.95
	西	41.400*	9.885	0.001	20.45	62.35
	南	50.200*	9.885	0.000	29.25	71.15

*均值差的显著性水平为 0.05。

从输出结果看，城东和城北、城西和城南之间的销售量没有显著的差异。

第三节　双因素方差分析

单因素方差分析只是考虑一个分类型自变量对数值型因变量的影响，但在实际的现实问题中，往往需要同时考虑多个影响因子对调查或试验结果的影响。如在分析影响考试分数的因素时，需要考虑授课教师、所在院系及学习时间等多因素的影响。为此，在进行多影响因子调查或试验并收集数据后，就可以采取多因子方差分析，从而得出各影响因子对所研究的变量的变化是否存在显著的影响。研究两个影响因子对结果的影响就是双因素方差分析。

一、双因素方差分析及类型

当方差分析中涉及一个因变量和两个分类型自变量时，称为一元双因素方差分析(由于本章只涉及一元方差分析，故后面将一元双因素方差分析简称为双因素方差分析)。

例 10-2：进行某项标准化考试，其分数在 200～800 分之间，分数越高越好。为尝试提高考生的分数，某校计划为学生提供考前辅导。学生主要来自于商学院、工学院和文学院，考前辅导老师有甲、乙、丙三人。

现按随机原则，从所有学生中采用分层抽样的方法，从商学院、工学院和文学院随机抽取了共 36 名学生分别参加了辅导前的测试，随后参加了三个老师的辅导，并统计了其辅导前、辅导后两次考试后的分数(如表 10-7 所示)，试图分析辅导教师和学生所在院系对考试成绩是否有显著的影响。

表 10-7　考试成绩

教师	商学院		工学院		文学院	
	考试成绩	辅导前成绩	考试成绩	辅导前成绩	考试成绩	辅导前成绩
教师甲	500.00	460.00	510.00	450.00	470.00	450.00
	520.00	480.00	520.00	460.00	490.00	460.00
	510.00	450.00	515.00	480.00	480.00	440.00
	530.00	460.00	530.00	490.00	500.00	400.00

续表

教师	商学院		工学院		文学院	
	考试成绩	辅导前成绩	考试成绩	辅导前成绩	考试成绩	辅导前成绩
教师乙	520.00	460.00	540.00	520.00	512.00	460.00
	560.00	500.00	550.00	530.00	500.00	440.00
	530.00	520.00	520.00	540.00	520.00	400.00
	570.00	510.00	580.00	450.00	510.00	460.00
教师丙	570.00	520.00	630.00	530.00	540.00	500.00
	620.00	500.00	610.00	540.00	580.00	510.00
	630.00	500.00	590.00	500.00	620.00	480.00
	670.00	520.00	660.00	540.00	590.00	490.00

上面问题中即属于双因素方差分析，若两影响因子之间相互独立，我们称之为无交互作用的双因素方差分析；若除两影响因子对因变量有影响外，两个影响因子的搭配对因变量也产生了新的影响效应，这时的方差分析被称为有交互作用的方差分析。

二、无交互作用的双因素方差分析

(一)数据结构

在无交互作用的双因素方差分析中，由于有两个影响因子，在获取时，需要将一个影响因子作为行因子(有 k 个水平)，另一个影响因子作为列因子(有 r 个水平)，行因子和列因子的每个水平都可以搭配成一组，观察不同情况下的因变量的取值。其数据结构如表 10-8 所示。

表 10-8 数据结构表

		列因子				平均值
		列因子水平 1	列因子水平 2	…	列因子水平 r	
行因子	行因子水平 1	x_{11}	x_{12}	…	x_{1r}	$\overline{x}_{1.}$
	行因子水平 2	x_{21}	x_{22}	…	x_{2r}	$\overline{x}_{2.}$
	…	…	…	…	…	…
	行因子水平 k	x_{k1}	x_{k2}	…	x_{kr}	$\overline{x}_{k.}$
平均值		$\overline{x}_{.1}$	$\overline{x}_{.2}$	…	$\overline{x}_{.r}$	$\overline{\overline{x}}$

(二)分析步骤

与单因素方差分析的步骤相似，双因素方差分析也是经过提出假设、构造检验统计量和做出判断等几个阶段。

(1)提出假设。由于要检验两个因素对因变量的影响，故需要对两个影响因子分别提出假设。

对行影响因子的假设：

$H_0: \mu_1 = \mu_2 \cdots = \mu_k$　　行因子对因变量无显著影响

$H_1: \mu_1, \mu_2, \cdots, \mu_k$ 不全相等　行因子对因变量有显著影响

对列影响因子的假设：

$H_0: \mu_1 = \mu_2 \cdots = \mu_r$　　列因子对因变量无显著影响

$H_1:\mu_1,\mu_2,\cdots,\mu_r$ 不全相等　列因子对因变量有显著影响

(2)构造并计算检验统计量。

按照方差分析的基本思想，需要将数据的差异进行分解：

总离差平方和 = 行因素产生的离差平方和
+ 列因素产生的离差平方和
+ 随机离差平方和

或表示为：　　$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSC}+\mathrm{SSE}$

或：　　$\sum_{i=1}^{k}\sum_{j=1}^{r}\left(x_{ij}-\bar{\bar{x}}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{i\cdot}-\bar{\bar{x}})^2+\sum_{i=1}^{k}\sum_{j=1}^{r}(\bar{x}_{\cdot j}-\bar{\bar{x}})^2+\sum_{i=1}^{k}\sum_{j=1}^{r}(x_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{\bar{x}})^2$

式中，$\bar{x}_{i\cdot}$ 为行因素第 i 个水平的样本均值;

$\bar{x}_{\cdot j}$ 为列因素第 j 个水平的样本均值。

在离差平方和计算的基础上，计算行因素的均方差、列因素的均方差和随机误差项的均方差。

行因素的均方差：　　$\mathrm{MSR}=\dfrac{\mathrm{SSR}}{k-1}$

列因素的均方差：　　$\mathrm{MSC}=\dfrac{\mathrm{SSC}}{r-1}$

随机误差项的均方差：　　$\mathrm{MSE}=\dfrac{\mathrm{SSE}}{(k-1)(r-1)}$

为检验行因素对因变量的影响是否显著，采用下面的统计量，当原假设成立时，该统计量服从自由度为 $k-1,(k-1)(r-1)$ 的 F 分布：

$$F_R=\frac{\mathrm{MSR}}{\mathrm{MSE}}\sim F(k-1,(k-1)(r-1))$$

同样，为检验列因素对因变量的影响是否显著，采用下面的统计量，当原假设成立时，该统计量服从自由度为 $r-1,(k-1)(r-1)$ 的 F 分布：

$$F_C=\frac{\mathrm{MSC}}{\mathrm{MSE}}\sim F(r-1,(k-1)(r-1))$$

(3)统计决策。

将我们上面所计算的检验统计量的值与根据显著性水平、自由度所确定的 F 分布的临界值进行比较，得出统计检验的结论并做出统计决策。

若 $F_R>F_\alpha$，则拒绝原假设，即 $H_0:\mu_1=\mu_2\cdots=\mu_k$ 不成立，行因素对因变量的影响显著;

若 $F_C>F_\alpha$，则拒绝原假设，即 $H_0:\mu_1=\mu_2\cdots=\mu_r$ 不成立，列因素对因变量的影响显著。

上述分析过程及结果的计算如表 10-9 所示。

表 10-9　分析过程及结果的计算

误差来源	误差平方和	自由度	均方差	F 值
	SS	df	MS	
行因素	SSR	k–1	MSR	F_R

续表

误差来源	误差平方和	自由度	均方差	*F* 值
	SS	df	MS	
列因素	SSC	$r-1$	MSC	F_C
随机误差	SSE	$(k-1)(r-1)$	MSE	
总和	SST	$kr-1$		

(二)关系强度的测量

关系强度测量的目的是要分析影响因子与因变量关系的密切程度，即影响因子对因变量变化的影响程度有多大。

由于我们将观测数据的总离差平方和分解为行因素产生的离差平方和、列因素产生的离差平方和和随机误差平方和三部分，即 $\mathrm{SST}=\mathrm{SSR}+\mathrm{SSC}+\mathrm{SSE}$。故要分析行因素和列因素对因变量影响的联合效应，可以计算行因素产生的离差平方和(也可称行因子效应)与列因素产生的离差平方和(也可称列因子效应)之和占数据总误差平方和也可称总效应的比重：

$$R^2=\frac{\text{联合效应}}{\text{总效应}}=\frac{\text{行因子效应}+\text{列因子效应}}{\text{总误差平方和}}=\frac{\mathrm{SSR}+\mathrm{SSC}}{\mathrm{SST}}$$

R^2 越大说明两个影响因子对因变量的影响程度越大，关系强度越大。

当然，我们也可以分别考察单个影响因子对变量的关系强度。

三、有交互作用的双因素方差分析

如果两个影响因子的搭配会对因变量产生新的影响作用，就需要考虑交互作用对因变量的影响，其数据结构如表 10-10 所示。

表 10-10　有交互作用方差分析的数据结构

		列因子			
		列因子水平 1	列因子水平 2	…	列因子水平 *r*
行因子	行因子水平 1	x_{111} x_{112} … x_{11m}	x_{121} x_{122} … x_{12m}	…	x_{1r1} x_{1r2} … x_{1rm}
	行因子水平 2	X_{211} X_{212} … X_{21m}	X_{221} X_{222} … X_{22m}	…	X_{2r1} X_{2r2} … X_{2rm}
	…	…	…	…	…
	行因子水平 *k*	X_{k11} X_{k12} … X_{k1m}	X_{k21} X_{k22} … X_{k2m}	…	X_{kr1} X_{kr2} … X_{krm}

上面数据结构中，两个影响因子的搭配组合重复观测 *m* 次。

其分析步骤如下。

(1)提出假设。

除去行因子和列因子的假设外，再加上交互作用对因变量影响的假设。

对行影响因子的假设：

$H_0:\mu_1=\mu_2\cdots=\mu_k$　　行因子对因变量无显著影响

$H_1:\mu_1,\mu_2,\cdots,\mu_k$不全相等　　行因子对因变量有显著影响

对列影响因子的假设：

$H_0:\mu_1=\mu_2\cdots=\mu_r$　　列因子对因变量无显著影响

$H_1:\mu_1,\mu_2,\cdots,\mu_r$不全相等　　列因子对因变量有显著影响

对交互作用的假设：

H_0：交互作用对因变量无显著影响

H_1：交互作用对因变量有显著影响

(2)进行数据差异的分解，构造并计算检验统计量。

总误差平方和 = 行因素产生的误差平方和
+ 列因素产生的误差平方和
+ 交互作用平方和
+ 随机误差平方和

或表示为：$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSC}+\mathrm{SSRC}+\mathrm{SSE}$

其中，总平方和为：$\mathrm{SST}=\sum_{i=1}^{k}\sum_{j=1}^{r}\sum_{l=1}^{m}(x_{ijl}-\bar{\bar{x}})^2$

行误差平方和为：$\mathrm{SSR}=rm\sum_{i=1}^{k}(\bar{x}_{i\cdot}-\bar{\bar{x}})^2$

列误差平方和为：$\mathrm{SSC}=km\sum_{i=1}^{k}(\bar{x}_{\cdot j}-\bar{\bar{x}})^2$

交互作用误差平方和为：$\mathrm{SSRC}=m\sum_{i=1}^{k}\sum_{=1}^{r}(\bar{x}_{ij}-\bar{x}_{i\cdot}-\bar{x}_{\cdot j}+\bar{\bar{x}})^2$

随机误差平方和为：$\mathrm{SSE}=\mathrm{SST}-\mathrm{SSR}-\mathrm{SSC}-\mathrm{SSRC}$

式中，x_{ijl}为对应行因素的第i个水平，列因素的第j个水平时的第l个观测值；

$\bar{x}_{i\cdot}$为行因素第i个水平的样本均值；

$\bar{x}_{\cdot j}$为列因素第j个水平的样本均值；

$\bar{x}_{ij}$为对应行因素的第i个水平，列因素的第j个水平组合的样本均值。

以此为基础，计算均方差。

行因素的均方差为：$\mathrm{MSR}=\dfrac{\mathrm{SSR}}{k-1}$

列因素的均方差为：$\mathrm{MSC}=\dfrac{\mathrm{SSC}}{r-1}$

交互作用的均方差为：$\mathrm{MSRC}=\dfrac{\mathrm{SSRC}}{(k-1)(r-1)}$

随机误差项的均方差为：$\mathrm{MSE}=\dfrac{\mathrm{SSE}}{kr(m-1)}$

最后计算检验统计量为：$F_R=\dfrac{\mathrm{MSR}}{\mathrm{MSE}}$，$F_C=\dfrac{\mathrm{MSC}}{\mathrm{MSE}}$，$F_{RC}=\dfrac{\mathrm{MSRC}}{\mathrm{MSE}}$

(3)将我们上面所计算的检验统计量的值与根据显著性水平、自由度所确定的 F 分布的临界值进行比较，得出统计检验的结论并做出统计决策。

四、双因素方差分析的软件操作与实现

利用 SPSS 软件进行多因子方差分析的相关计算，其操作方法及计算结果如下。

SPSS 选项：【Analyze】—【General Linear Model】—【Univariate】。

第一步：将变量“考试成绩”放入【Dependent Variables】中。

第二步：将变量“教师”和“所在学院”放入【Fixed Factor】中。

第三步：单击 Options 按钮进入，选择【Homogemety of Variance Test】(进行方差分析的假定条件：所有总体的研究变量的方差相等的检验)项，选择【Means Plots】(画均值图)项。

第四步：单击【Post Hoc】进入，选择【Equal Variances Assumed】中的“LSD”(进行多重比较检验)。

第五步：单击【OK】按钮。

计算结果如表 10-11 所示。

表 10-11 计算结果输出表

因变量：考试成绩

Source	差的平方和	自由度	均方差	F 值	P 值
模型	78 347.000(a)	8	9793.375	16.504	0.000
截距	10 886 700.250	1	10 886 700.250	18 346.367	0.000
教师	67 922.167	2	33 961.083	57.232	0.000
所在学院	10 322.167	2	5161.083	8.698	0.001
教师 * 所在学院	102.667	4	25.667	.043	0.996
随机误差	16 021.750	27	593.398		
合计	10 981 069.000	36			

从输出结果看，两个因子的显著性概率都小于 1%。所以在 1%的显著性水平下，认为两个因子都是显著性因子。即教师和学生所在学院对考试成绩都有显著的影响。但是教师与学生所在学院的交互作用概率值为 0.996，大于 5%的显著性水平，因此教师与学生所在学院之间的交互作用对考试成绩来说不存在显著影响。

多因子方差分析的应用前提与单因子方差分析一样，也需要满足方差一致性(方差相等)，从对方差一致性检验的结果看(表 10-12)，Sig=0.254>0.05，因此不能拒绝方差相同的原假设，即该检验结果显示满足方差一致性的要求。

表 10-12 对方差一致性检验的结果

F 值	第一自由度	第二自由度	P 值
1.370	8	27	.254

五、包含协变量的多因子方差分析

从前面所进行的方差分析过程及结果中，有人可能会注意到在进行方差分析时，变量“辅导前的测试成绩”并没有对作为考试成绩的因子进行分析。而辅导前的测试成绩对最终的考试成绩是否会有影响呢？为什么没有将该变量放入控制变量(Factor)中进行分析呢？这是因为在方差分析中要求控制变量是可控的，即放入控制变量中的变量必须是定性变量。但实际中，有些因子的不同水平很难人为控制，但他们确确实实对观测变量产生了显著的影响。如辅导前的测试成绩可能对考试成绩会有影响。在方差分析中如果忽略这些因子的存在，而单单去分析其他因子对观测变量的影响，往往会夸大或缩小这些因子的影响作用，使得分析结论不正确。

因此为了更加准确地研究控制变量的不同水平对因变量的影响，应尽量排除其他能够排除的因子对分析的影响，即将那些很难控制的因子(定量变量)作为协变量，进行方差分析。引入协变量的方差分析称为协方差分析。

协方差分析是将那些很难控制的数值型的影响因素作为协变量，在排除协变量影响的条件下，分析控制变量对因变量的影响，从而更加准确地对控制因子的作用进行评价。

方差分析中的因子都是定性变量，而协方差分析中的协变量应是定量的、数值型的变量，即连续数值型。协变量之间没有交互影响，且与控制变量之间也没有交互影响。

这时的方差分析模型即包含协变量的方差分析的模型为：

因变量=因素 A 主效应+因素 B 的主效应+因素 A 与 B 的交互效应+协变量+随机误差

进行斜方差分析的方法与方差分析一样，只不过要将变量“辅导前的测试成绩”放入 Covariatie 中即可。

再看方差分析的结果(如表 10-13 所示)，协变量“辅导前的成绩”对最终考试成绩是没有显著的影响的，而教师和所在学院两个因子对考试成绩的影响是显著的。因此本例中可以不考虑协变量的影响。

表 10-13　因子效应检验(Tests of Between-Subjects Effects)

Source	差的平方和	自由度	均方差	F 值	P 值
模型	78 636.815	9	8737.424	14.440	0.000
截距	24 408.154	1	24 408.154	40.339	0.000
辅导前成绩	289.815	1	289.815	0.479	0.495
教师	37 603.417	2	18 801.709	31.073	0.000
所在学院	7588.652	2	3794.326	6.271	0.006
教师 * 所在学院	120.698	4	30.175	0.050	0.995
随机误差	15 731.935	26	605.074		
合计	10 981 069.000	36			

此外，我们也可以从因子“教师”的不同水平均值比较检验结果中看出(如表 10-14 所示)，不同教师的辅导对考试成绩存在显著影响，即对考试成绩产生了显著差异。

来自不同院的学生的考试成绩又如何呢？表 10-15 表明商学院与工学院的学生在考试成绩上没有显著的差异，而文学院学生的考试成绩明显低于另外两个学院学生的成绩。

最后，我们从(表 10-16)“教师”与“所在学院”的交互分析表中可知，教师丙的平均成绩最高，其次是教师乙，考试成绩最低的是教师甲。在每个教师的辅导下，商学院与工学院的学生的平均成绩基本一致，而文学院的成绩都比较低。

表 10-14 成对比较(Pairwise Comparisons)

(I) 教师	(J) 教师	均值差(I-J)	标准误	P 值	95% 差的置信区间	
					下限	上限
教师甲	教师乙	−28.083(*)	9.945	0.009	−48.488	−7.678
	教师丙	−102.917(*)	9.945	0.000	−123.322	−82.512
教师乙	教师甲	28.083(*)	9.945	0.009	7.678	48.488
	教师丙	−74.833(*)	9.945	0.000	−95.238	−54.428
教师丙	教师甲	102.917(*)	9.945	0.000	82.512	123.322
	教师乙	74.833(*)	9.945	0.000	54.428	95.238

表 10-15 成对比较(Pairwise Comparisons)

(I) 所在学院	(J) 所在学院	均值差(I-J)	标准误	P 值	95% 差的置信区间	
					下限	上限
商学院	工学院	−2.083	9.945	0.836	−22.488	18.322
	文学院	34.833(*)	9.945	0.002	14.428	55.238
工学院	商学院	2.083	9.945	0.836	−18.322	22.488
	文学院	36.917(*)	9.945	0.001	16.512	57.322
文学院	商学院	−34.833(*)	9.945	0.002	−55.238	−14.428
	工学院	−36.917(*)	9.945	0.001	−57.322	−16.512

表 10-16 教师所在学院

辅导教师	所在学院	Mean(均值)	Std.Error (标准误)	95% 置信区间	
				下限	上限
教师甲	商学院	515.000	12.180	490.009	539.991
	工学院	518.750	12.180	493.759	543.741
	文学院	485.000	12.180	460.009	509.991
教师乙	商学院	545.000	12.180	520.009	569.991
	工学院	547.500	12.180	522.509	572.491
	文学院	510.500	12.180	485.509	535.491
教师丙	商学院	622.500	12.180	597.509	647.491
	工学院	622.500	12.180	597.509	647.491
	文学院	582.500	12.180	557.509	607.491

思考与练习

1．如何理解方差分析的原理？

2．方差分析的作用是什么？

3．进行方差分析的前提条件有哪些？如何验证这些条件是否满足？

4．什么是方差分析中的交互效应？如何判断交互效应是否显著？

5．什么是协变量？为什么要进行协方差分析？

6．在五个地区中随机抽取了几天的发生交通事故的次数如下：

交通事故样本数据

东部	北部	中部	南部	西部
15	12	10	14	13
17	10	14	9	12
14	13	13	7	9
11	17	15	10	14
—	—	—	7	9

由于是随机抽样，有一些地区的样本容量较多(如南部和西部)，而有些地区的样本容量较少(如东部)。试以 α=0.01 的显著水平检验各地区平均每天发生交通事故的次数是否相等。

第十一章　列联分析与对应分析

第一节　交叉分组与列联表

一、交叉分组

在实际分析中我们常常要研究两个或多个分类型变量之间的关系，如性别与收入等级之间、学历与职业选择偏好之间、居住地区与收视偏好之间的关系等。随着人们生活水平的提高，工作压力的加大及家庭的需要，每个人对家庭中的女性是否应该就业的观点不尽相同。有人提出了应该“男人在外工作，妇女在家操持家务”的就业观点，这个观点是否代表大多数人的意见，特别是妇女们本身是否同意该观点，它与妇女的文化程度有何关系等是我们需要研究的问题。

某地区针对此问题对已婚妇女进行了抽样调查，我们在调查资料中关注了文化程度和就业观点两个变量，并对被调查者按其“文化程度”和对“男人在外工作，妇女在家操持家务”的就业观点的态度的两个变量同时进行分组，并按照两变量各自的分类水平列在一张表中，计数后编制成下表(如表 11-1 所示)。

表 11-1　按文化程度分的妇女就业观点

文化程度	就业观点			
	1-非常同意	2-同意	3-不同意	4-非常不同意
1-小学及以下	8	82	96	11
2-初中	41	220	327	48
3-高中	72	224	503	47
4-大学	24	61	300	41

这种将所有个体单位同时按照两个变量进行分组，并形成交叉分组表的方法称为交叉分组。

二、列联表

所谓列联表就是由两个或两个以上的变量进行交叉分类的频数分布表[表 11-1 就是一个列联表(Contingency table)]，对列联表的分析就称为列联分析。

在列联表中，每个变量都有两个或更多的可能取值，这些取值也称为水平，如文化程度有四个水平、就业观点有四个水平等。

一般将列联表中横向变量的划分类别数记为 R，纵向变量的划分类别数记为 C，称这样的列连表为 $R\times C$ 列联表，上表即为 4×4 列联表。

列联表一般包括观察值分布和期望值分布两种。

(一)观察值分布

事实上，表 11-1 就是一个最简单的观察值的分布。观察值分布反映了数据的实际分布，但总体的基数(即规模)不同时，分布的特征并不适合于直接用表中的频数进行对比。为

了能进行更有效的比较分析，在列联表中得到更多的有价值的信息，往往需要计算相应的百分比，即得到频率分布，如表 11-2 所示。

表 11-2 文化程度·就业观点 Cross Tabulation（交叉制表）

文化程度		就业观点				Total
		非常同意	同意	不同意	非常不同意	
小学及以下	Count 频数	8	82	96	11	197
	% within 文化程度	4.1%	41.6%	48.7%	5.6%	100.0%
初中	Count 频数	41	220	327	48	636
	% within 文化程度	6.4%	34.6%	51.4%	7.5%	100.0%
高中	Count 频数	72	224	503	47	846
	% within 文化程度	8.5%	26.5%	59.5%	5.6%	100.0%
大学及以上	Count 频数	24	61	300	41	426
	% within 文化程度	5.6%	14.3%	70.4%	9.6%	100.0%
Total	Count 频数	145	587	1226	147	2105
	% within 文化程度	6.9%	27.9%	58.2%	7.0%	100.0%

从频率分布表中我们可以对所分析对象有一些初步的认识：如回答“同意”和“不同意”的人的比例最高，随着文化程度的提高，就业观点持“同意”的比例下降，持“不同意”的比例增多。但是妇女的文化程度与就业观点是存在有一定联系还是彼此独立，仅从百分比上很难得出结论，这需要我们进一步分析。

（二）期望值分布

所谓期望分布是指在某种假设成立的情况下的理论分布。如我们假设妇女的文化程度与就业观点之间是独立的，则不同文化程度的妇女的观点应该是相同的，即不同文化程度的妇女在对“男人在外工作，妇女在家操持家务”这一观点的“非常同意”、“同意”、“不同意”、“非常不同意”的比率是相同的。从调查结果看，所有被调查的妇女回答“非常同意”的人数占 6.9%。如果两变量之间相互独立，则各种学历妇女回答“非常同意”的人数比重均为 6.9%，相应的人数如下。

小学及以下学历的妇女回答“非常同意”的人数应为：197×6.9%=14 人。

初中学历妇女回答“非常同意”的人数应为：636×6.9%=44 人。

高中学历妇女回答“非常同意”的人数应为：846×6.9%=58 人。

大学学历妇女回答“非常同意”的人数应为：426×6.9%=29 人。

这 14 人、44 人、58 人和 29 人就是本例中的期望频数，以此类推可以计算出其他就业观点期望值的分布，如表 11-3 所示。

表 11-3 期望值分布

文化程度	就业观点			
	1-非常同意	2-同意	3-不同意	4-非常不同意
1-小学及以下	14	55	115	14
2-初中	44	177	370	45
3-高中	58	236	492	59
4-大学	29	119	248	30

我们也可以把观察值与期望值的频数分布列在一张表中(如表 11-4 所示)。

表 11-4　观察值与期望值频数对比分布表

文化程度		就业观点			
		1 非常同意	2 同意	3 不同意	4 非常不同意
1-小学及以下	观察值	8	82	96	11
	期望值	14	55	115	14
2-初中	观察值	41	220	327	48
	期望值	44	177	370	45
3-高中	观察值	72	224	503	47
	期望值	58	236	492	59
4-大学	观察值	24	61	300	41
	期望值	29	119	248	30

第二节　变量独立性的检验与相关测量

如果假设两变量是独立的，则实际观察到的数据频数分布与“不同文化程度的妇女对该就业观点看法相同”的假设下的理论的期望分布应该很相近。从表 11-4 中可以看到观察值分布和期望值分布还是有差异的，但这种差异如果不是很大，完全可以由样本的随机性解释，就说明一定程度上反映出实际情况与假设很可能是一致的。然而，要判断在总体上是否是一致的，我们需要进行统计检验。

若要检验两个分类型变量是否独立，或者说两者是否存在一定的相关，我们需要采用 χ^2 检验的方法。

一、χ^2 检验统计量

用 f_0 表示观察值分布中的频数，f_e 表示期望值分布中的频数，则 χ^2 统计量为：

$$\chi^2=\sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(f_0-f_e)^2}{f_e}$$

如果不同文化程度的人员对该就业观点的看法不同，即就业观点受文化程度影响，那么表 11-4 中观察值和期望值频数分布就会相差较大，相反如果两变量之间是独立的，则实际观测分布与理论期望分布应基本一致。如果两变量之间独立，则上面 χ^2 统计量服从 χ^2 分布。

由于 χ^2 值的大小与观察值和期望值的配对数 $R\times C$ 有关，所以 χ^2 统计量的分布与自由度 $(R-1)(C-1)$ 有关。

χ^2 统计量的曲线图随其自由度不同显示出不同的图形，如图 11-1 所示。

根据 χ^2 统计量我们可以计算出 χ^2 值，再利用 χ^2 分布计算出 χ^2 值对应的概率，然后我们就可以据此做出是否拒绝原假设的判断。

二、变量独立性检验

和一般的统计检验相似，两个分类型变量独立性的检验也需要提出原假设、构造并计算检验统计量、根据显著性水平进行判断得出统计结论等几个步骤。下面以上面调查的问题为例，对“文化程度”与“就业观点”两个变量之间的独立性进行检验。

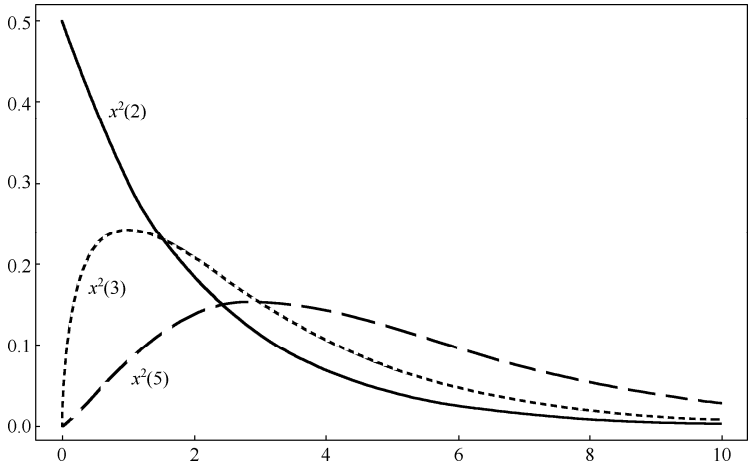


图 11-1 χ^2 统计量的曲线图

(1)提出假设，即：

$H_0: P_{\text{小学及以下}} = P_{\text{初中}} = P_{\text{高中}} = P_{\text{大学}}$ 文化程度与就业观点无关

H_1：$P_{\text{小学及以下}}$、$P_{\text{初中}}$、$P_{\text{高中}}$、$P_{\text{大学}}$ 不全相等 文化程度与就业观点有关系

构造检验统计量：

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{4}\frac{(f_0 - f_e)^2}{f_e} = \frac{(8-14)^2}{14} + \frac{(82-55)^2}{55} + \cdots + \frac{(41-30)^2}{30} = 85.761$$

该统计量服从自由度为 $(R-1)\times(C-1) = (4-1)\times(4-1) = 9$ 的 χ^2 分布。

(2)根据事先确定的显著性水平 α 及自由度，查到 χ^2 拒绝域的临界值，将 χ^2 检验统计量的值与临界值比较，做出拒绝原假设的结论，或根据计算得到的 P 值与显著性水平 α 比较，得到同样的结果。

对于这个假设的统计检验，我们也可以利用统计软件实现。

卡方检验的 SPSS 实现：

SPSS 选项：【Analyze】—【Descriptive Statistics】—【Crosstabs】。

第一步：将变量“文化程度”放入【Rows】中，变量“就业观点”放入【Columns】中。

第二步：单击【statistics】按钮，选择 χ^2 检验。

第三步：单击【OK】按钮即可。

χ^2 检验结果如表 11-5 所示。

表 11-5 Chi-Square Tests(卡方检验)

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square(χ^2 检验统计量)	85.761(a)	9	0.000
Likelihood Ratio(似然比)	89.135	9	0.000
Linear-by-Linear Association(线性联合)	26.806	1	0.000
N of Valid Cases(有效样本)	2105		

a 0 cells (.0%) have expected count less than 5. The minimum expected count is 13.57.

从输出结果(如表 11-5 所示)看，卡方值=85.761，对应的概率值=0.000，是小概率，所以我们可以拒绝原假设，即认为不同文化程度的妇女对该就业观点的看法不同，两者之间存在关系。

当然，实际应用中有不止一个 χ^2 检验统计量，除此以外还包括似然比(Likelihood Ratio)检验统计量和线性联合分布(Linear-by-Linear Association)统计量。它们都有渐近的 χ^2 分布，虽然计算的结果不同，但对应的概率值都相同，结论是一致的。

三、列联表中变量的相关度测量

对列联表中变量之间的相互关系进行检验，如果拒绝原假设，则认为变量之间存在联系，那么接下来的问题就是它们之间的相关性有多大？

回答此问题时，我们要根据所分析的变量类型来选择不同的相关检验方法。SPSS 软件中提供了多种相关检验的方法。

Interval by interval：定距变量与定距变量相关的检验。

Nominal by nominal：名义变量与名义变量相关的检验。

Ordinal by ordinal：顺序变量与顺序变量相关的检验。

列联表中两变量相关程度测量的 SPSS 实现:

SPSS 选项:【Analyze】—【Descriptive Statistics】—【Crosstabs】。

下一步：选择【statistics】项，再根据所分析的变量类型进行选择即可。

由于我们分析的数据都是顺序变量，所以选择【Ordinal by ordinal】的方法，详见图 11-2，在【Ordinal by ordinal】方法中有四种检验方法，选择哪一种都可以。

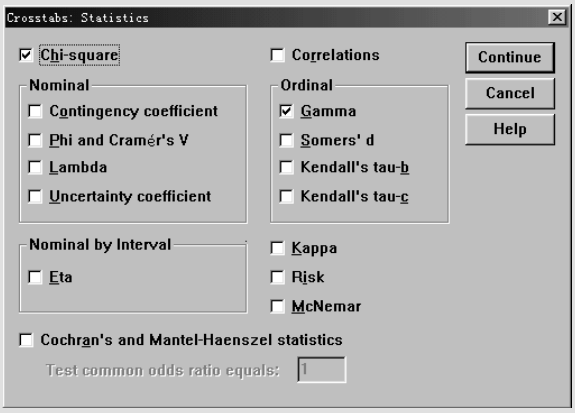


图 11-2

相关关系的输出结果如表 11-6 所示。

表 11-6　相关检验的输出结果

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Ordinal by Ordinal	Gamma	0.190	0.029	6.413	0.000
N of Valid Cases		2105			

a　Not assuming the null hypothesis.

b　Using the asymptotic standard error assuming the null hypothesis.

输出结果说明妇女文化程度与就业观点的相关关系显著。

四、应用中的准则

利用 χ^2 分布进行检验时，要求样本容量必须足够大，特别是每个单元中的期望频数不能过小，否则应用 χ^2 检验可能会得出错误的结论。一般地，进行 χ^2 检验有以下两个基本准则：

第一，如果只有两行两列，每个单元格的期望频数必须是 5 或以上。

第二，如果有两个以上的行或列，则期望频数小于 5 的单元格数目低于 20%，否则不能应用 χ^2 检验。

例如，表 11-7 中的数据可以进行卡方检验，因为 6 个单元中只有 1 个单元格的期望频数小于 5，但是表 11-8 中的数据不能应用卡方检验。

如果我们仔细观察会发现，表 11-8 中的 f_0 与 f_e 非常接近，最大的差别只是 3，应当说期望值与观察值基本一致，它们之间并无显著差别。然而，用 χ^2 检验得到的结果却是拒绝原假设，差异显著。最简单的解决方法是：将小单元合并，使得 f_e 大于 5。

表 11-7　频数分布表

类别	f_0	f_e
A	28	26
B	49	47
C	18	23
D	6	4
E	92	88
F	20	25
合计	213	213

表 11-8　频数分布表

类别	f_0	f_e
A	30	32
B	110	113
C	86	87
D	23	24
E	5	2
F	5	4
G	4	1
合计	263	263

第三节　对应分析[①]

通过列联分析我们对交叉汇总表的数据进行了分析，了解到妇女对“男人在外工作，妇女在家操持家务”的就业观点与其文化程度有关，并且有显著的相互关系，即文化程度不同的妇女对该就业观点的看法是不相同的。但是从列联分析中无法直观而简单地给出各分类之间的联系，我们希望在此数据基础上将文化程度与就业观点的关系直观地表现在一张二维图中(如图 11-3 所示)，从而反映他们之间的对应关系，这种分析方法就是对应分析。

这就是我们前面所说的二维对应分析图。我们可以从对应分析图中看到，各种受教育程度按高低次序分列于横轴两端，说明不同文化程度的妇女，在就业观点上有明显的差异。具体在两个变量的关系上，我们看到具有大学文化程度的妇女与不同意和非常不同意接近，而具有高中及以下文化程度的妇女与同意和非常同意接近，说明受教育程度较高的妇女不同意“男人在外工作，妇女在家操持家务”的就业观点，而中等及以下文化程度的妇女却持赞同观点。

① 建议学习本节内容前先学习因子分析。

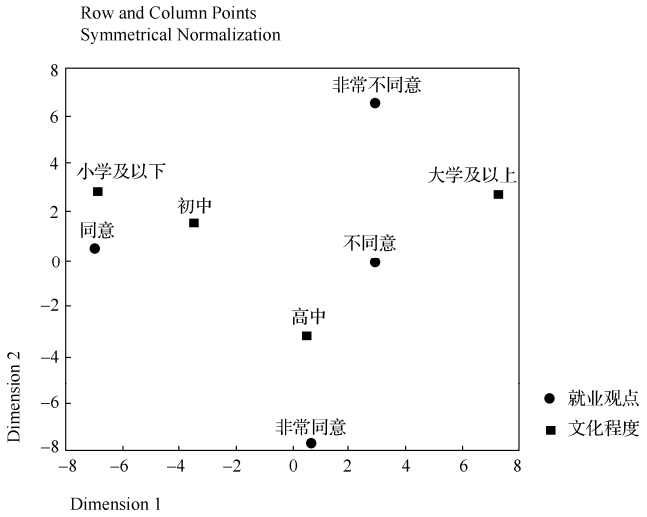


图 11-3　文化程度与就业观点的关系图

一、对应分析的原理

对应分析是多元统计分析方法，它利用降维的思想以达到简化数据结构的目的。不过与因子分析不同的是，它同时对数据表中的行与列进行处理，寻求以低维图形(如二维图)表示数据表中行变量与列变量的关系。对应分析方法广泛用于对由定性变量构成的列联表数据的研究，利用对应分析可以在一张二维图上同时画出属性变量不同取值或类别的情况，以直观、简洁的形式描述属性变量各种状态之间的相互关系，以及不同属性变量之间的相互关系。同时，对应分析也可应用于定量变量数据的分析，揭示变量与样品之间的相互关系。

对应分析是将指标型的因子分析与样品型的因子分析结合起来进行的统计分析。它是从指标型因子分析出发，从而直接获得样品因子分析的结果。概括起来，对应分析可以提供以下三方面的信息：

(1)指标之间的关系；

(2)样品之间的关系；

(3)指标与样品之间的关系。

由于指标型的因子分析与样品型的因子分析都是反映一个整体的不同侧面，因此它们之间一定存在内在联系。对应分析就是通过一个过渡矩阵 $\boldsymbol{Z}$ 将二者有机地结合起来。

二、对应分析计算机实现与输出结果解读

SPSS 中对应分析的操作如下。

SPSS 选项：【Analyze】—【Data Reduction】—【Correspondence Analysis】。

第一步：将“文化程度”变量选入【Rowk】中，单击下方的【Define Range】按钮，填入其取值范围 1～4，并单击右侧的【update】按钮；把“就业观点”选入【Column】中，在其【Define Range】中输入取值范围 1～4，并单击右侧的【update】按钮。

第二步：如果想分别得到对应分析中行变量和列变量的散点图，可在【Plots】中的【Scatter plot】里单击【Row points】(行变量图)项和【Column points】(列变量图)项。单击【OK】按钮，即可得到输出结果。

SPSS 输出结果如下。

(1)各维汇总表如表 11-9 所示。

表 11-9 各维汇总表

维度	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation
1	0.187	0.035			0.859	0.859	0.020	0.037
2	0.070	0.005			0.122	0.981	0.022	
3	0.028	0.001			0.019	1.000		
Total		0.041	85.761	0.000(a)	1.000	1.000		

a 9 degrees of freedom

为了更好地理解输出表中数据的含义，现介绍对应分析表中出现的一些基本概念。

Singular Value——奇异值(是惯量的平方根)，反映了行与列各水平在二维图中分量的相关程度，是对行与列进行因子分析产生的新的综合变量的典型相关系数。

Inertia——惯量，实际上就是常说的特征根，表示的是每个维度对各个变量类别之间差异的解释量。

Chi Square——卡方，就是关于列联表行列独立性检验的χ^2 统计量的值，和前面表中的相同。其后面的 Sig 为在行列独立的零假设下的 P 值，注释表明自由度为(4–1)×(3–1)=6，Sig 值很小说明列联表的行与列之间有较强的相关性。

Proportion of Inertia——惯量比例，是各维度(公因子)分别解释总惯量的比例及累计百分比，类似于因子分析中公因子解释能力的说明。

从上面的表 11-9 中，我们可以确定需要使用多少个维度对结果进行解释。第一列 dimension 为维度，第二列 Singular Value 翻译为奇异值，是惯量(Inertia)的平方根，实际上它相当于相关分析中的相关系数，而第三列 Inertia 也就是惯量，就是特征根，用于说明对应分析各个维度的结果能够解释列联表中两个变量联系的程度。上表可见第一维特征根值为 0.187，第二维为 0.070。右侧 Proportion of Inertia 是各维度特征根的解释程度，第一维能解释全部信息的 85.9%，第二维解释 12.2%，前两维的累计解释程度为 98.1%。因此，二维图形可以表达出 98.1%的原始数据信息。而且，观察时以第一维度为主即可。

(2)行变量-文化程度的输出结果如表 11-10 所示。

表 11-10 行变量-文化程度的输出结果(a)

文化程度	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
小学及以下	0.094	–0.684	0.279	0.009	0.234	0.104	0.896	0.056	0.952
初中	0.302	–0.344	0.152	0.008	0.192	0.099	0.893	0.066	0.958
高中	0.402	0.053	–0.318	0.003	0.006	0.578	0.069	0.928	0.997
大学及以上	0.202	0.725	0.276	0.021	0.569	0.219	0.948	0.052	1.000
Active Total	1.000			0.041	1.000	1.000			

a Symmetrical normalization

表 11-10 说明了行变量的有关内容。

第一部分是关于行变量每一类别在两个维度上的分值情况，实际上就是每一类别在坐标图(如图 11-3 所示)中的坐标，Mass 为行与列的边缘概率。

第二部分(Contribution of Point to Inertia of Dimension)是说明行变量各个类别对每一维度特征值的影响，数值越大的类别，说明它对类别间差异的影响越大。

第三部分(Contribution of Dimension to Inertia of Point)是说明每一维度对行变量各个类别特征值的影响。最后一列是说明第一和第二维度对每一个变量总共可以解释的百分比。

从表中看，除了高中文化以外，其他各文化程度类别特征值的分布都以第一维度为主，说明各类别间的差异绝大部分都反映在第一维度中。并且，所有文化程度可由第一和第二维度解释的比例都在95%以上，说明通过该二维表可以解释原变量的绝大部分信息。

(3)列变量-就业观点的输出结果如表11-11所示。

表 11-11　列变量-就业观点的输出结果(a)

就业观点	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
非常同意	0.069	0.060	–0.766	0.003	0.001	0.573	0.015	0.899	0.914
同意	0.279	–0.690	0.046	0.025	0.709	0.008	0.998	0.002	1.000
不同意	0.582	0.287	–0.009	0.009	0.257	0.001	0.986	0.000	0.987
非常不同意	0.070	0.297	0.649	0.004	0.033	0.418	0.320	0.577	0.898
Active Total	1.000			0.041	1.000	1.000			

a　Symmetrical normalization

表 11-11 输出的是列变量的有关内容，其形式与前面行变量分析表一样，含义的解释也与行变量的相同，这里不再赘述。

从表 11-11 中看到，同意与不同意的特征值分布以第一维度为主，即其差异反映在第一维度上；非常同意与非常不同意的特征值分布主要以第二维度为主。由于第一维度解释了大部分信息，所以同意与不同意的差异是主要的。此外，除非常不同意以外，其他就业观点可由第一和第二维度解释的比例都在90%以上，说明通过该二维表可以解释原变量的绝大部分信息。

(4)行变量文化程度各个类别之间的分值分布图。

图 11-4 表现的是行变量文化程度各个类别之间的分值分布，从中可看出各个类别间的差异。由图 11-4 可见，小学及以下和初中类别间的差异不大(横向看)，可归为一类，他们与大学的差别最大。

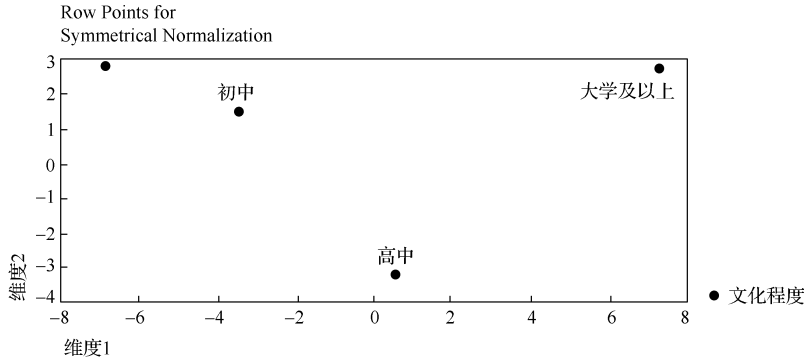


图 11-4　行变量文化程度各个类别之间的分值分布图

(5)列变量文化程度各个类别之间的分值分布图。

图 11-5 表现的是列变量各个类别之间的分值分布，从中可看出各个类别间的差异。由

图 11-5 可见，从横轴上看，同意与不同意的差别最大，而非常不同意和非常同意的差异主要体现在纵轴上。

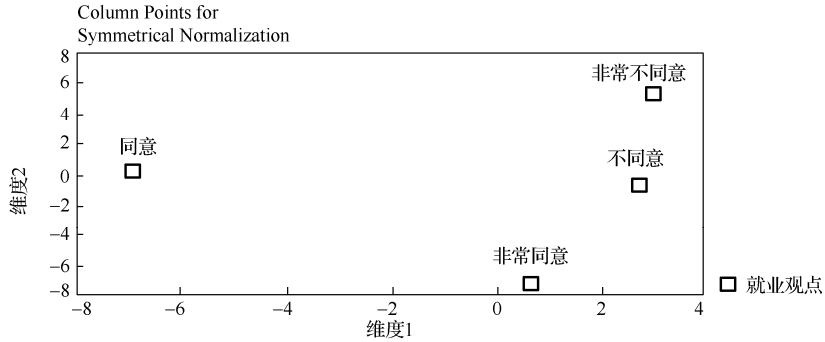


图 11-5 列变量文化程度各个类别之间的分值分布图

(6)两个变量的二维对应分析图。

该图是对应分析输出结果中最重要的部分，即我们前面所说的二维对应分析图。实际上，很多情况下我们只需看该图就可以知道分析结果了。我们从以下两个方面阅读该图形：

首先分别检查两个变量在横轴(第一维度)和纵轴(第二维度)上的区分情况，如果同一变量的某些不同类别在某个方向上比较接近，则说明这些类别在该维度上区别不大。

其次，要比较两个变量各个取值类别间的位置关系，在同一个方向上靠近的两种类别其联系的程度就更大些。因此，图 11-6 直观地反映了行变量与列变量之间的联系，从中可看出对该观点持赞同态度最高的是高中文化程度，其次为初中、小学和小学以下，而具有大学及以上文化程度的人持有不赞同、非常不赞同的观点。

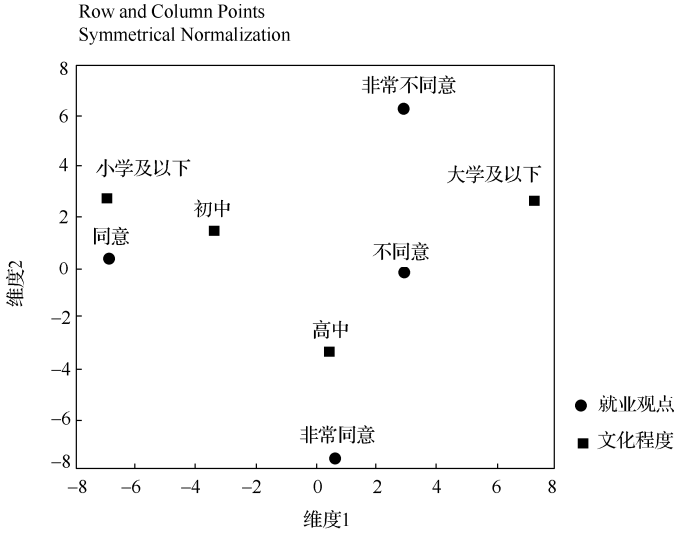


图 11-6 两个变量的二维对应分析图

三、对应分析应用——失业原因与教育程度关系分析

目前我国面临着前所未有的就业压力，由于各种原因也造成了各式各样的失业情况，我们现在想研究的问题是失业与受教育程度有何关系？

城镇失业人员的失业原因有不同的情况，可分为下岗离开单位，毕业后未找到工作，原单位破产，辞职、被辞退或合同期满及其他五种情况；受教育程度可分为不识字、小学、初中、高中、大专和大学本科及以上。我们现在所关心的问题是，这些失业原因受哪些因素的影响？或者在一个更具体的分析中，我们关心这些失业原因与失业人员的受教育水平之间究竟有没有什么联系？它们之间的关系是怎样的？

我们可以将失业原因与失业人员的受教育水平按照各自的分类水平划出一张交叉汇总表，如表 11-12 所示。

表 11-12　按受教育程度分的城镇失业人员失业原因构成(%)

受教育程度	1 下岗离开单位	2 毕业后未找到工作	3 原单位破产	4 辞职、被辞退或合同期满	5 其　他
不识字	22.0	2.4	17.1	7.3	51.2
小学	39.0	9.5	18.6	6.8	26.1
初中	45.2	18.1	13.3	9.7	13.8
高中	43.5	23.3	11.7	13.0	8.6
大专	30.1	35.8	11.1	16.7	6.5
大学本科及以上	19.4	43.7	11.7	18.4	6.8

数据来源：《中国统计年鉴 2003》

表 11-12 给出了按受教育程度来区分的城镇失业人员失业原因构成，但从该表中无法直观而简单地给出各分类之间的联系，解释起来也略显复杂。首先可以先做一下卡方检验，观察两者之间是独立的还是有相互关系的。

卡方检验的结果如表 11-13 所示。

表 11-13　卡方检验的结果

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	167.365(a)	20	0.000
Likelihood Ratio	164.012	20	0.000
Linear-by-Linear Association	29.098	1	0.000
N of Valid Cases	602		

a　0 cells (.0%) have expected count less than 5. The minimum expected count is 11.84.

卡方检验结果表明，失业原因与受教育程度之间不独立，是有相互关系的。因此我们可以进行相关分析。由于失业原因是分类变量，而受教育程度是顺序变量，所以我们选择了 Nominal by nominal(名义变量与名义变量相关的检验)与 Ordinal by ordinal(顺序变量与顺序变量相关的检验)两种检验方法(如表 11-14 所示)。

表 11-14　检验计算结果

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	Contingency Coefficient	.466			.000
Ordinal by Ordinal	Gamma	−.160	.040	−3.921	.000
N of Valid Cases		602			

a　Not assuming the null hypothesis.

b　Using the asymptotic standard error assuming the null hypothesis.

从检验结果看，不论哪种方法其近似概率值都为 0.000，表明了各类变量之间相关关系显著，但是从相关系数来看，两种检验结果的值不同，而且符号相反。那么失业原因与受教育程度之间到底是一种怎样的关系，这从列联分析中是得不到明确答案的，这也是列联分析的局限性。因此下面我们接着做对应分析。通过对应分析将失业原因与受教育程度的关系可以直观地表现在一张二维图中。

对应分析 SPSS 的输出结果如下。

(1) 各维汇总表如表 11-15 所示。

表 11-15 各维汇总表

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	0.477	0.227			0.812	0.812	0.034	0.314
2	0.225	0.050			0.180	0.992	0.039	
3	0.045	0.002			0.007	1.000		
4	0.011	0.000			0.000	1.000		
Total		0.280	168.055	0.000(a)	1.000	1.000		

a 20 degrees of freedom

从各维汇总表中，我们可以确定需要使用多少个维度对结果进行解释。由表 11-5 可见第一维特征根值为 0.227，第二维特征根值为 0.050。右侧 Proportion of Inertia 是各维度特征根的解释程度，第一维能解释全部信息的 81.2%，第二维能解释 18%，前两维的累计解释程度为 99.2%。因此，二维图形可以表达出 99.2%的原始数据信息。而且，观察时以第一维度为主即可。

(2) 行变量——教育程度的输出结果如表 11-16 所示。

表 11-16 行变量——教育程度的输出结果

教育程度	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
不识字	0.167	–1.243	0.507	0.133	0.540	0.191	0.926	0.073	0.998
小学	0.167	–0.491	–0.343	0.025	0.084	0.087	0.774	0.178	0.952
初中	0.167	0.036	–0.548	0.011	0.000	0.223	0.009	0.981	0.990
高中	0.167	0.304	–0.453	0.015	0.032	0.152	0.476	0.499	0.974
大专	0.167	0.622	0.177	0.032	0.135	0.023	0.961	0.037	0.998
本科	0.167	0.771	0.661	0.064	0.208	0.324	0.741	0.256	0.997
Active Total	1.000			0.280	1.000	1.000			

表 11-16 说明了行变量的有关内容。

从表中看，除了初中和高中文化以外，其他各文化程度类别特征值的分布都以第一维度为主，说明各类别间的差异绝大部分都反映在第一维度中。并且，所有文化程度可由第一和第二维度解释的比例都在 95%以上，说明通过该二维表可以解释原变量的绝大部分信息。

(3) 列变量——失业原因的输出结果如表 11-17 所示。

表 11-17　列变量——失业原因的输出结果

失业原因	Mass	Score in Dimension		Inertia	Contribution				
					Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
		1	2		1	2	1	2	Total
下岗	0.332	0.021	–0.638	0.031	0.000	0.601	0.002	0.995	0.997
毕业未工作	0.221	0.885	0.425	0.092	0.363	0.178	0.901	0.098	0.999
单位破产	0.139	–0.262	–0.032	0.006	0.020	0.001	0.755	0.005	0.760
辞职	0.120	0.480	0.327	0.016	0.058	0.057	0.800	0.175	0.975
其他	0.188	–1.189	0.441	0.135	0.558	0.163	0.938	0.061	0.999
Active Total	1.000			0.280	1.000	1.000			

从表 11-17 中可以看出，除失业原因为下岗以外，其他各原因的特征值分布都以第一维度为主，即其差异主要反映在第一维度上；只有下岗的特征值分布主要以第二维度为主。此外，所有失业原因中除单位破产外，可由第一和第二维度解释的比例都在 97%以上，说明通过该二维表可以解释原变量的绝大部分信息。

(4) 行变量文化程度各个类别之间的分值分布图如图 11-7 所示。

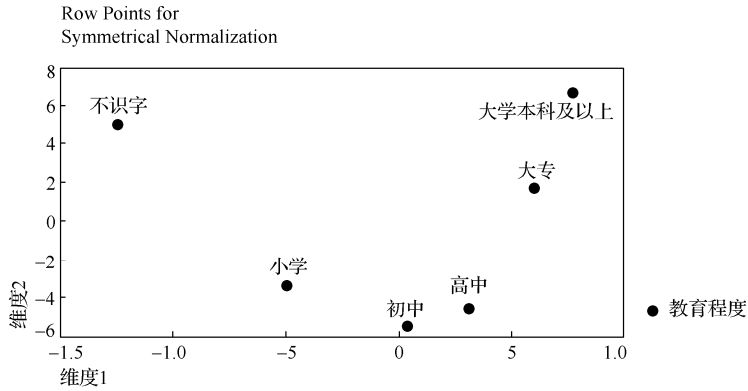


图 11-7　行变量文化程度各个类别之间的分值分布图

(5) 列变量失业原因各个类别之间的分值分布图如图 11-8 所示。

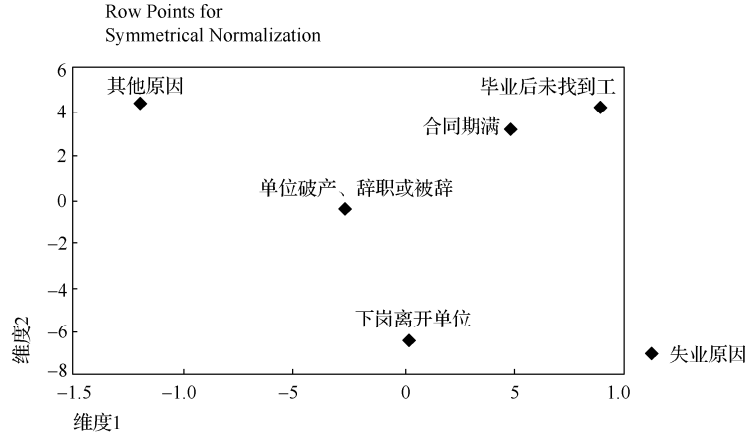


图 11-8　列变量失业原因各个类别之间的分值分布图

(6)两个变量的二维对应分析图如图 11-9 所示。

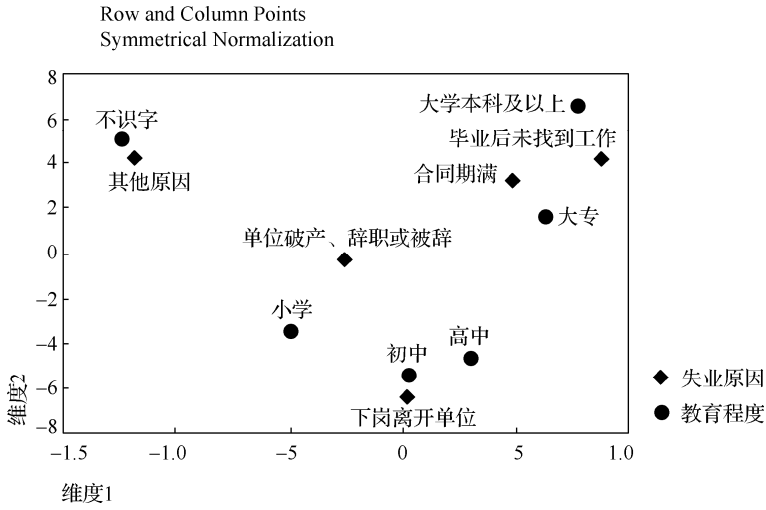


图 11-9 两个变量的二维对应分析图

从对应分析图中我们看到，教育程度和失业原因在第一维度上有很大的距离，而坐标轴的刻度在第二维度上区分不太明显，这和前面说的数据变异主要以第一维为主的结论一致。各受教育程度按高低次序分列于横轴(即第一维度)两端，说明不同教育程度的职工在失业原因上有明显的差异。具体从两个变量的关系上，我们看到本科、大专两个类别与失业原因变量中的辞职、被辞退或合同期满，毕业未找到工作之间的关系接近，说明这两类受教育程度较高的人群在失业原因上自主性较强，小学、初中、高中主要与下岗及原单位破产之间关系接近，在失业原因上比较被动。而不识字和其他失业原因关系接近。

思考与练习

1．简述列联分析的主要内容和基本原理。

2．对各高校各专业学生进行抽样调查，了解其性别和生源地，编制列联表，并检验专业与性别是否有关。

3．什么是对应分析？其应用对象是什么？

4．对应分析方法的思想是什么？

5．对某年北京市居民收入水平和消费支出情况进行对应分析。

6．尝试就某问题进行问卷调查，并利用调查的数据资料进行列联分析和对应分析。

7．利用下面数据资料对某地区机关人员的学历与对发展西部地区急需解决的问题看法进行对应分析(数据资料见下表)。

学历与对西部地区急需解决的问题看法数据资料

	资金	人才	新观念	信息	相关政策	知识与技术
高中及以下	48	15	7	2	2	5
大专	68	27	29	2	6	3
本科	16	15	16	2	4	1

8．对受教育程度与失业原因进行对应分析(数据资料见下表)。

按受教育程度分的城镇失业人员失业原因构成(%)

受教育程度	下岗离开单位(a)	毕业后未找到工作(b)	原单位破产(c)	辞职、被辞退或合同期满(d)	其他(e)
不识字	22.0	2.4	17.1	7.3	51.2
小学	39.0	9.5	18.6	6.8	26.1
初中	45.2	18.1	13.3	9.7	13.8
高中	43.5	23.3	11.7	13.0	8.6
大专	30.1	35.8	11.1	16.7	6.5
大学本科	19.4	43.7	11.7	18.4	6.8

第十二章　相关与回归分析

社会、经济与自然现象之间的相关联系和制约是一个普遍的现象。社会经济的发展可以表现为经济变量的数量方面，因而社会经济现象内部和外部各种相互的联系就表现为经济变量间的关联。

本章重点研究的是数值型变量与数值型变量之间的关系，通过对变量之间关系的分析与研究，为预测和科学决策提供依据。如企业销售部经理在深入了解了广告费用和销售收入之间的关系后，才能尝试去预测一定水平的广告费用支出可能带来多少销售收入，从而做出科学的决策。又例如，企业在了解销售利润率与产品产量、产品价格等因素之间的关系后，才可以通过各影响因素的变化去预测可能的销售利润率水平，并以此作为企业经营决策的依据。在研究居民消费支出的问题时，人们通常认为，提高收入水平会带动消费支出水平的提高。这一认识我们也可以从表 12-1 中的数据资料及相关图(图 12-1)中可以看到，人均可支配收入水平越高的地区，其消费支出水平也相对较高；相反人均可支配收入水平较低的地区，其消费支出水平也相对较低。进一步地，我们想知道，怎么判断两个变量之间是否相关，它们的相关形式是什么？相关的程度如何？我们是否可以通过收入水平去估计、预测消费支出的水平？要想回答这些问题，就需要进行变量之间的相关关系分析及回归分析。这正是本章要解决的主要问题。

表 12-1　某年各省市居民人均可支配收入与人均消费支出(元)

地　　区	居民人均可支配收入	居民人均消费支出	地　　区	居民人均可支配收入	居民人均消费支出
北京市	44 488.57	31 102.89	湖北省	18 283.23	12 928.31
天津市	28 832.29	22 342.98	湖南省	17 621.74	13 288.73
河北省	16 647.4	11 931.54	广东省	25 684.96	19 205.5
山西省	16 538.32	10 863.83	广　西	15 557.08	10 274.31
内蒙古	20 559.34	16 258.12	海南省	17 476.46	12 470.59
辽宁省	22 820.15	16 067.98	重庆市	18 351.9	13 810.62
吉林省	17 520.39	13 025.97	四川省	15 749.01	12 368.4
黑龙江	17 404.39	12 768.76	贵州省	12 371.06	9303.35
上海市	45 965.83	33 064.76	云南省	13 772.21	9869.54
江苏省	27 172.77	19 163.56	西　藏	10 730.22	7316.95
浙江省	32 657.57	22 551.97	陕西省	15 836.75	12 203.59
安徽省	16 795.52	11 726.99	甘肃省	12 184.71	9874.57
福建省	23 330.85	17 644.47	青海省	14 373.98	12 604.8
江西省	16 734.17	11 088.89	宁　夏	15 906.78	12 484.52
山东省	20 864.21	13 328.9	新　疆	15 096.62	11 903.71
河南省	15 695.18	11 000.44			

现象之间数量关系的研究，统计上可以从两个方面进行：一方面是分析现象之间是否有关系及其关系的密切程度；另一方面是找出现象之间的数量依存关系。本章针对此问题介绍相关与回归分析的基本理论与方法，包括相关分析与回归分析两个基本方法。

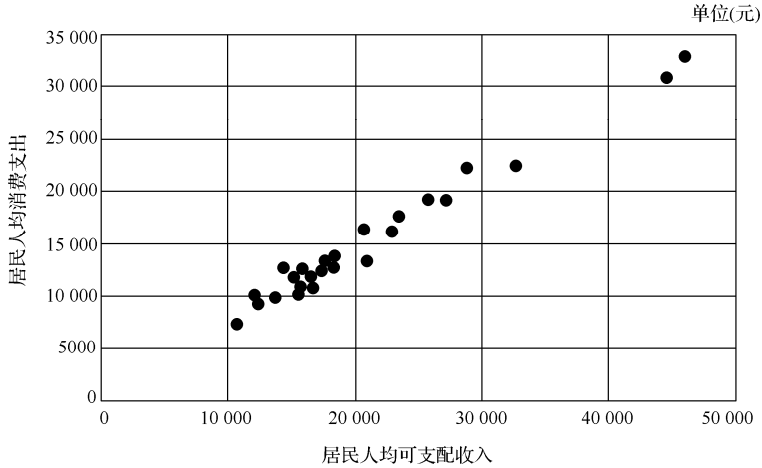


图 12-1　2014 年居民人均可支配收入与人均消费支出相关图

第一节　相关分析

一、相关关系

各种现象之间、变量之间相互作用、相互联系的方式各不相同，其关系的密切程度也不尽相同。各种关系中，一种极端的情况是一个变量的变化可以完全决定另一个变量的变化。例如，银行的一年期的存款利率是 2.55%，若存入的本金用 x 表示，则一年后的本利和为 $y=x+2.55\%x$，在此本利和与本金之间是一种确定性的函数关系，在利率不变的情况下，本金的大小可以完全决定一年期的本利和。虽然经济现象、自然现象中还有很多类似的函数关系，但还有更多的情况是事物之间有密切的联系，然而它们的密切程度并没有达到一个现象可以完全决定另一个现象的程度，如总产出与投资额之间、粮食产量与施肥量之间、广告费用支出与产品销售额之间等。此外，上面提到的人均可支配收入与消费支出之间存在着密切的关系，收入水平提高了，消费支出就增加，但是由于消费者支出水平不仅受收入水平的影响，还受到价格、消费习惯、年龄、收入预期等很多因素的影响，因而可支配收入并不能完全决定消费支出水平，这样两者之间就是一种非确定的关系。我们将这种关系，即存在着一定的联系但又不是严格的、确定的关系称为相关关系，本章所要研究的也正是这种相关关系。

相关分析主要是分析现象之间是否存在相关关系，以及相关关系的方向、形式和关系的密切程度。具体来说，相关分析的主要内容有以下几方面。

(1) 确定现象之间有无关系。这是相关分析的起点，只有存在一定的关系，才有必要进行进一步的分析。

(2) 确定相关关系的表现形式。只有判明了现象之间相关关系的具体表现形式，才能进一步地运用相应的回归分析方法去分析，研究变量之间的数量依存关系。如果把非线性相关误认为是线性相关，进而确定线性回归方程，便会出现认识上的偏差，导致错误的结论。

(3) 测定相关关系的密切程度。现象之间的相关关系是一种不严格的数量关系，因此给人的感觉是松散的。相关分析就是要从这种松散的数量关系中，判定其相关关系的密切程度。

二、相关关系的描述——散点图

对于两个变量 X 和 Y，通过观察和实验，我们可以得到其若干组数据，记为 (x_i, y_i) $(i=1,2,\cdots,n)$，将这些数据按 x 的值由大到小(或由小到大)以序列表表示，即构成相关表。通过相关表可以粗略地看出两个变量之间存在着相关关系，如果两者之间是变化方向是一致的，即存在着正相关；而若两者之间的变化方向是相反的，则为负相关，如将表 12-1 中数据按居民人均可支配收入从高到低进行排序形成下面相关表(如表 12-2 所示)。

表 12-2 某年居民人均可支配收入与人均消费支出相关表 单位：元

地区	居民人均可支配收入	居民人均消费支出	地区	居民人均可支配收入	居民人均消费支出
上海市	45 965.83	33 064.76	安徽省	16 795.52	11 726.99
北京市	44 488.57	31 102.89	江西省	16 734.17	11 088.89
浙江省	32 657.57	22 551.97	河北省	16 647.4	11 931.54
天津市	28 832.29	22 342.98	山西省	16 538.32	10 863.83
江苏省	27 172.77	19 163.56	宁　夏	15 906.78	12 484.52
广东省	25 684.96	19 205.5	陕西省	15 836.75	12 203.59
福建省	23 330.85	17 644.47	四川省	15 749.01	12 368.4
辽宁省	22 820.15	16 067.98	河南省	15 695.18	11 000.44
山东省	20 864.21	13 328.9	广　西	15 557.08	10 274.31
内蒙古	20 559.34	16 258.12	新　疆	15 096.62	11 903.71
重庆市	18 351.9	13 810.62	青海省	14 373.98	12 604.8
湖北省	18 283.23	12 928.31	云南省	13 772.21	9869.54
湖南省	17 621.74	13 288.73	贵州省	12 371.06	9303.35
吉林省	17 520.39	13 025.97	甘肃省	12 184.71	9874.57
海南省	17 476.46	12 470.59	西　藏	10 730.22	7316.95
黑龙江	17 404.39	12 768.76			

如果将一一对应的两变量观测值 (x_i, y_i) 描点于坐标系上，即构成散点图，又称为相关图，如图 12-1 所示，从散点图我们可以看到人均可支配收入越高，人均消费支出总体上也表现出越多的特点，甚至从散点图我们也可以大致看到这两个变量存在什么形式的相关关系。

绘制两变量的相关图的 SPSS 实现：

SPSS 选项：【Graphs】—【Scatter】—在 SPSS 中的提供的四种散点图(简单散点图、重叠散点图、矩阵散点图和三维散点图)中选定散点图的类型，如简单散点图(Simple)—【Define】—指定散点图 y 轴上的变量名到【y Axis】框中，指定散点图的 x 轴上的变量名到【x Axis】框中—单击【OK】按钮，即可。

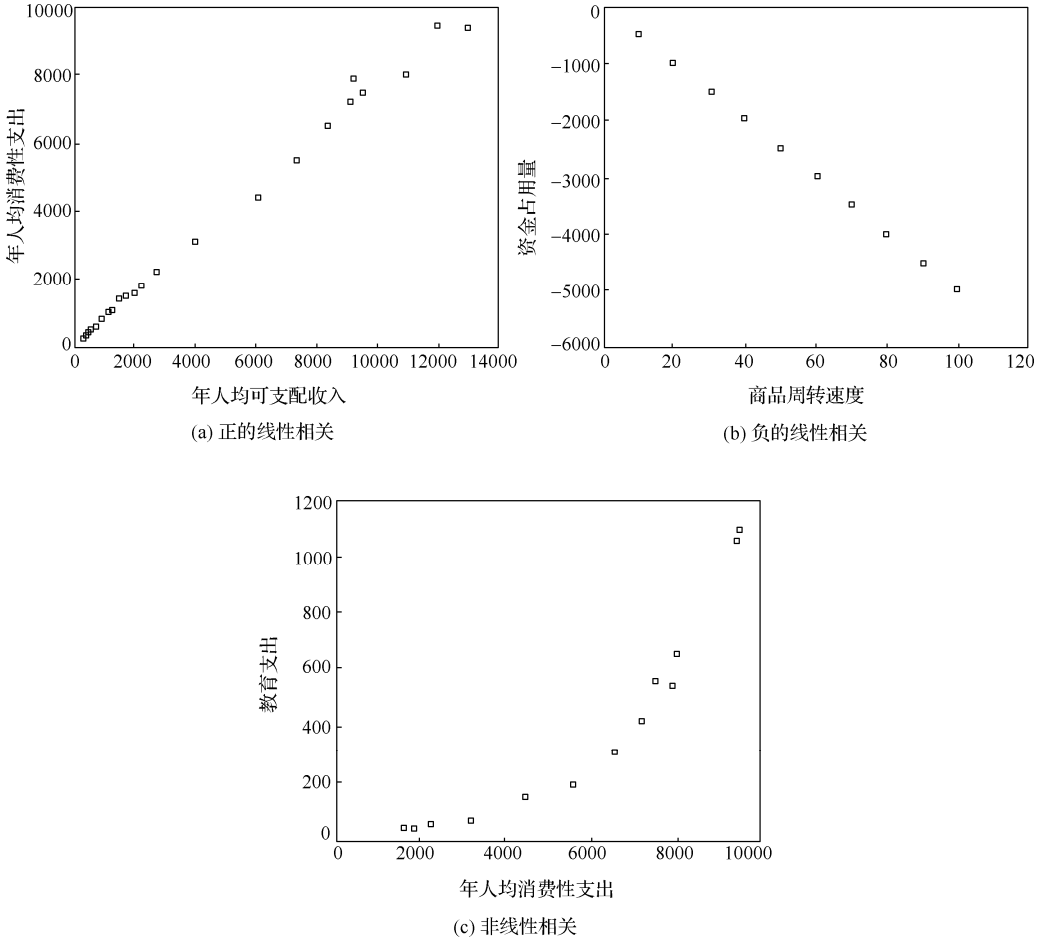
通过相关图所反映出的坐标点的分布状况，可以更直观地判断变量之间是否存在相关关系，以及相关的形态、方向。

(1)相关的形态。若变量 Y 与变量 X 的相关关系表现为线性组合，或绘制的散点图近似地表现为一条直线，则称之为线性相关，如图 12-2(a)和图 12-2(b)所示；若 Y 与 X 是非线性组合，或绘制的散点图近似地表现为一条曲线，则称之为非线性相关或曲线相关，如图 12-2(c)所示。

(2)相关的方向。当两个变量的变动方向相同，即一个变量增加，另一个变量总体上也

表现出相应地增加的态势，或一个变量减少，另一个变量总体上也相应地减少时，两个变量之间的关系称为正相关，如图 12-2(a)所示；若两个变量变动的方向相反，即一个变量增加的同时，另一个变量随之减少，两个变量之间的关系则称为负相关，如图 12-2(b)所示。

通过图 12-1 可以看到，人均可支配收入与消费支出之间属于线性相关、正相关。



(a) 正的线性相关　(b) 负的线性相关

(c) 非线性相关

图 12-2　相关散点图

三、相关程度的测定——相关系数的计算

散点图虽然能够直观地展现变量之间的相关关系，但很不精确。相关系数则是测定变量之间关系密切程度的量，它能够以定量的方式准确地描述变量之间的相关程度。相关系数有多种，对于不同类型的变量数据，应计算不同的相关系数。

(一) Pearson 简单相关系数

1. Pearson 简单相关系数的计算

Pearson 简单相关系数是最常用的相关系数，它是用来度量两个定量变量 X 与 Y 之间的线性相关程度，如人均可支配收入与消费支出的相关程度、身高与体重之间的相关程度等。利用总体数据计算的相关系数称为总体相关系数，一般用 ρ 表示。而很多情况下，我们所掌握的只是样本

数据，利用样本数据计算的相关系数称为样本相关系数，用 r 表示，其计算公式是：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

Pearson 简单相关系数的基本原理是把每一对观测值 (x_i,y_i) 中的 x_i 值与均值 $\bar{x}$ 的距离与相应的 y_i 值与均值 $\bar{y}$ 的距离相乘，如果这个乘积为正，则说明相对于各自的均值，两个变量的变化趋势一样；如果这个乘积为负，那么说明他们的变化趋势相反。把样本中所有这些乘积相加，如果样本中的乘积多为正，则和为正；如果样本中的乘积多为负，则和为负。如果乘积正负号的个数差不多，则乘积和就接近零，再将其标准化就得到上面的相关系数的计算公式。

2．Pearson 简单相关系数的性质与其具体含义

r 的取值范围在–1～1 之间，即–1≤ r ≤1。$r>0$ 表明两个变量之间存在正线性相关关系；$r<0$ 表明两个变量之间存在负线性相关关系；当 $|r|=1$ 时，表示为完全相关；当 r=0 时，表现为无线性相关；当 $0<|r|<1$ 时，表现为不完全相关。

r 具有对称性，即 X 与 Y 之间的相关系数与 Y 与 X 之间的相关系数相等。

r 的数值的大小与 X 和 Y 的计量尺度无关，改变 X 和 Y 的数据的计量尺度，并不改变 r 的数值。

r 是两个变量之间线性关系的度量指标，但无法反映两变量之间的因果关系，即使 r 很高，也不一定意味着 X 与 Y 之间一定存在着因果关系。

此外，应该注意的是，Pearson 简单相关系数是反映两个变量的线性相关程度，但它并不能够度量变量之间的非线性相关程度。

3．总体相关的显著性检验

由于样本的随机性，样本数量少等原因，利用样本数据计算出来的相关系数不能直接说明总体变量之间是否存在显著相关，需要进行统计检验。检验的方法与步骤如下。

首先，确定原假设 H_0:两变量之间不存在线性相关，或 ρ=0。

其次，计算检验统计量 $t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$，该统计量服从 n–2 个自由度的 t 分布。

第三，查 t 分布表，看 t 检验统计量是否落在拒绝域中，若是则拒绝原假设，可以认为两变量之间存在显著的相关关系；或利用统计软件直接得到伴随概率，如果伴随概率小于或等于指定的显著性水平，则拒绝原假设，可以认为两变量之间不存在显著的相关关系，否则不能拒绝原假设。

（二）Spearman 等级相关系数

Spearman 等级相关系数用来度量顺序变量间的线性相关程度，它不能直接通过变量值计算，而是根据数据的秩(即按样本数据大小排序的位次)进行计算的，它适合有序数据或不满足正态分布假设的等间隔数据。

Spearman 等级相关系数的计算公式是：

$$\theta=\frac{\sum(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum(R_i-\bar{R})^2\sum(S_i-\bar{S})^2}}$$

式中，R_i 是第 i 个 X 值的秩；S_i 第 i 个 Y 值的秩；$\overline{R}$、$\overline{S}$ 分别是 R_i 和 S_i 的平均值。

同样，利用样本数据计算了 Spearman 等级相关系数后，应需要对总体两个变量的等级相关的显著性进行统计检验，原假设为不存在显著的等级相关，一般如果样本数小于或等于 30 时，则利用 Spearman 等级相关统计量表(具体的检验统计量、相关概念及计算可以参考非参数统计方面的教科书)，SPSS 将自动依据此表给出对应的伴随概率值；如果样本量大于 30，SPSS 将计算 Z 检验统计量：$Z = R\sqrt{n-1}$，该统计量近似服从正态分布，SPSS 将根据正态分布表给出相应的伴随概率值。

(三)Kendall tau-b 等级相关系数

Kendall tau-b 等级相关系数也是一种利用变量的秩通过非参数统计反映两个有序变量或两个秩变量间的相关程度的分析方法。其度量的基本原理是把所有的样本点配对，然后看每一对中的 x 和 y 的观测值是否同时增加或减少。若两个变量同时增长或下降，称这两点协同，否则就是不协同。如果样本中协同的点数目多，两个变量就相关，如果不协同的多，两变量就不相关(具体的检验统计量、相关概念及计算可以参考非参数统计方面的教科书)。

Kendall tau-b 等级相关系数的计算，我们可以用 SPSS 计算相关系数。对 Kendall tau-b 等级相关系数的检验时，一般若样本量小于等于 30，则直接利用 Kendall 等级相关统计表，SPSS 将自动给出对应的伴随概率，若样本量大于 30，则计算检验统计量：$Z = \dfrac{3T\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$，该检验统计量服从标准正态分布，SPSS 将直接给出相应的伴随概率。

上面三个相关系数取值均在–1～1 之间，其值越接近 1(或–1)正(或负)相关程度越高，越接近 0，相关程度越低。然而到底如何算接近，不能一概而论，需要进行统计检验，其中后两个相关系数的检验属于非参数检验的范畴(与总体分布无关)。

相关系数计算及检验的 SPSS 实现：

(1)SPSS 选项：【Statistics】—【Correlate】—【Bivariate】。

(2)在【Correlation Coefficents】框中，选择计算的相关系数。

(3)在【Test of Significance】框中，确定统计检验是双尾检验(Tow-tailed)还是单尾检验(One-railed)，一般若并未明确两变量间是正相关还是负相关时，可以做双尾检验，否则可以选择单尾检验。

利用表 12-1 中的数据资料计算的 Pearson 简单相关系数结果如表 12-3 所示。

表 12-3　Pearson 简单相关系数表

		人均可支配收入	人均消费支出
人均可支配收入	Pearson 相关性	1	0.989**
	显著性(双侧)		0.000
	N	31	31
人均消费支出	Pearson 相关性	0.989**	1
	显著性(双侧)	0.000	
	N	31	31

**. 在 0.01 水平(双侧)上显著相关。

表 12-3 显示了人均可配收入与人均消费支出之间的 Pearson 相关系数。结果是以对角

线形式给出的，每个单元格共分三行，分别是相关系数、*P* 值(Sig)和样本数。可以看到人均消费支出与人均可支配收入的相关系数为 0.989，对相关系数检验双侧的 *P* 值小于 0.01，所以可以认为两者之间存在有显著的且高度相关关系。

利用表 12-1 中的数据资料计算出的 Kendall tau-b 与 Spearman 等级相关系数结果如表 12-4 所示。

表 12-4 Kendall tau-b 与 Spearman 等级相关系数

			人均可支配收入	人均消费支出
Kendall 的 tau-b	人均可支配收入	相关系数	1.000	0.819**
		Sig(双侧)	—	0.000
		N	31	31
	人均消费支出	相关系数	0.819**	1.000
		Sig(双侧)	.000	—
		N	31	31
Spearman 的 rho	人均可支配收入	相关系数	1.000	0.935**
		Sig(双侧)	—	0.000
		N	31	31
	人均消费支出	相关系数	0.935**	1.000
		Sig(双侧)	0.000	—
		N	31	31

**. 在置信度(双侧)为 0.01 时，相关性是显著的。

表 12-4 显示的是人均可支配收入与人均消费支出的等级相关系数，表明两变量之间的等级相关程度较高。

(四)偏相关系数

偏相关系数描述的是在控制了一个或几个变量的条件下两个变量之间的相关程度，如若控制年龄和学历两个因素，分析消费支出与人均可支配收入之间的相关程度；在控制年龄和工作经历两个因素的条件下分析工资收入与受教育程度之间的相关程度；在控制价格因素的条件下分析广告费用与销售量之间的相关程度等。

在控制变量 z 的情况下，计算变量 x 与 y 之间的偏相关系数的计算公式为：

$$r_{xy,z}=\frac{r_{xy}-r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

控制了变量 z_1、z_2，变量 x 与 y 之间的偏相关系数的计算公式为：

$$r_{xy,z_1z_2}=\frac{r_{xy,z_1}-r_{xz_2,z_1}r_{yz_2,z_1}}{\sqrt{(1-r_{xz_2,z_1}^2)(1-r_{yz_2,z_1}^2)}}$$

偏相关系数检验的统计量是 $t=\dfrac{\sqrt{n-k-2}\cdot r}{\sqrt{1-r^2}}$，其中 r 是相应的偏相关系数。

偏相关系数计算及检验的 SPSS 实现：

(1)SPSS 选项：【Statistics】—【Correlate】—【Partial】。

(2)在【Varible】框中选择需计算相关系数的两个变量，在【Contorling for】框中，选择需要控制的变量即可。

城镇居民人均可支配收入、人均消费支出和 GDP 的数据资料如表 12-5 所示。

表 12-5　某年居民人均可支配收入与人均消费支出数据表

地区	居民人均可支配收入(元)	居民人均消费支出(元)	国内生产总值(亿元)	地区	居民人均可支配收入(元)	居民人均消费支出(元)	国内生产总值(亿元)
上海市	45 965.83	33 064.76	23 567.7	安徽省	16 795.52	11 726.99	20 848.75
北京市	44 488.57	31 102.89	21 330.83	江西省	16 734.17	11 088.89	15 714.63
浙江省	32 657.57	22 551.97	40 173.03	河北省	16 647.4	11 931.54	29 421.15
天津市	28 832.29	22 342.98	15 726.93	山西省	16 538.32	10 863.83	12 761.49
江苏省	27 172.77	19 163.56	65 088.32	宁　夏	15 906.78	12 484.52	2752.1
广东省	25 684.96	19 205.5	67 809.85	陕西省	15 836.75	12 203.59	17 689.94
福建省	23 330.85	17 644.47	24 055.76	四川省	15 749.01	12 368.4	28 536.66
辽宁省	22 820.15	16 067.98	28 626.58	河南省	15 695.18	11 000.44	34 938.24
山东省	20 864.21	13 328.9	59 426.59	广　西	15 557.08	10 274.31	15 672.89
内蒙古	20 559.34	16 258.12	17 770.19	新　疆	15 096.62	11 903.71	9273.46
重庆市	18 351.9	13 810.62	14 262.6	青海省	14 373.98	12 604.8	2303.32
湖北省	18 283.23	12 928.31	27 379.22	云南省	13 772.21	9869.54	12 814.59
湖南省	17 621.74	13 288.73	27 037.32	贵州省	12 371.06	9303.35	9266.39
吉林省	17 520.39	13 025.97	13 803.14	甘肃省	12 184.71	9874.57	6836.82
海南省	17 476.46	12 470.59	3500.72	西　藏	10 730.22	7316.95	920.83
黑龙江省	17 404.39	12 768.76	15 039.38				

若直接计算人均消费支出和地区生产总值之间的相关系数，可以得到两者之间的相关系数为 0.336(如表 12-6(a)所示)。但若将人均可支配收入作为控制变量，计算人均消费支出与地区生产总值的偏相关系数，则可以看到两者之间的相关系数为–0.290[如表 12-6(b)所示]，其主要原因是因为人均可支配收入与地区生产总值和人均消费支出之间存在一定的相关，若将人均可支配收入固定不变，则人均消费支出与地区生产总值之间相关系数发生了变化。

表 12-6(a)　相关系数

		人均消费支出	国内生产总值
人均消费支出	Pearson 相关性	1	0.336
	显著性(双侧)	—	0.065
	N	31	31
国内生产总值	Pearson 相关性	0.336	1
	显著性(双侧)	0.065	—
	N	31	31

表 12-6(b)　偏相关系数

控制变量			国内生产总值	人均消费支出
人均可支配收入	国内生产总值	相关性	1.000	–0.290
		显著性(双侧)	—	0.121
		d*f*	0	28
	人均消费支出	相关性	–0.290	1.000
		显著性(双侧)	0.121	—
		d*f*	28	0

第二节 线性回归分析

相关分析旨在测度变量之间关系的密切程度，它所使用的测量工具就是相关系数。而回归分析则是考察若干自变量 $X_1,X_2,\cdots,X_p$ 与因变量 Y 之间的数量依存关系的统计方法和技术。回归分析的内容主要包括有以下几个方面。

(1) 从样本数据出发，确定变量之间数量依存关系的数学关系式，即回归模型的形式。

(2) 估计回归模型的参数。

(3) 对所确定的、估计的回归模型的可信程度进行各种统计检验，并从影响因变量的诸多变量中找出影响显著的自变量。

(4) 利用回归模型，根据一个或几个自变量的值来预测或控制因变量的水平，并给出相应的精确度。

一、线性回归模型

(一) 理论回归模型

描述变量 Y 与 $X_1,X_2,\cdots,X_p$ 之间线性关系的数学结构式，即线性理论回归模型为：

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\varepsilon$$

一般，我们称 Y 为被解释变量(因变量)，称 $X_1,X_2,\cdots,X_p$ 为解释变量(自变量)，β_i 为未知参数，其中 β_0 为回归常数，$\beta_1,\beta_2,\cdots,\beta_p$ 为回归系数。

该模型显示出 Y 与 $X_1,X_2,\cdots,X_p$ 之间的关系可以用两个部分描述：一部分是由于 $X_1,X_2,\cdots,X_p$ 的变化引起的 Y 的变化的部分，即 $\beta_0+\beta_1X_1+\cdots+\beta_pX_p$；另一部分是由除去 $X_1,X_2,\cdots,X_p$ 外的其他一切被忽略和无法考虑到的随机因素引起的，即 ε，我们称其为随机误差项。

建立线性回归模型时，需要假定被解释变量 Y 与解释变量 $X_1,X_2,\cdots,X_p$ 之间具有线性关系，而 $X_1,X_2,\cdots,X_p$ 之间不存在高度的线性关系，且解释变量的取值是非随机的(即其值是外生的、事先给定的)，被解释变量则是随机变量。这就意味着，对于给定的解释变量 $X_1,X_2,\cdots,X_p$ 的值，Y 的取值都相应地对应着一个分布。此外，对于随机误差项 ε 需要做出以下假定。

(1) 正态性，ε 是一个服从正态分布的随机变量。

(2) ε 的期望值为 0，即：$E(\varepsilon)=0$。

这样，对 $X_1,X_2,\cdots,X_p$ 两边求数学期望得：

$$E(Y)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p$$

该式从平均意义上表达了变量 Y 与 $X_1,X_2,\cdots,X_p$ 的统计规律性，关于这一点，在应用上是非常重要的，因为我们通常关心的正是这个平均值。

(3) 方差齐性，对于任何一组特定的 $X_1,X_2,\cdots,X_p$ 值，ε 的方差 σ^2 都相同。

(4) 独立性，对于任何一组特定的 $X_1,X_2,\cdots,X_p$ 值，它所对应的 ε 与其他一组 $X_1,X_2,\cdots,X_p$ 所对应的 ε 不相关。这样，对于该特定的 $X_1,X_2,\cdots,X_p$ 值，它所对应的 Y 值与其他 $X_1,X_2,\cdots,X_p$ 所对应的 Y 值也不相关。在解释变量 $X_1,X_2,\cdots,X_p$ 值一定的情况下，Y 的

变化由误差项 ε 的方差 σ^2 来决定。当 σ^2 较小时，Y 的实际观测值与估计值就比较接近；当 σ^2 较大时，Y 的实际观测值与估计值偏离就比较大。

（二）一元线性回归模型与多元线性回归模型

在回归模型 $Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\varepsilon$ 中，当 $p=1$ 时，即影响被解释变量的主要因素只有一个，模型中引入一个解释变量时，该模型被称为一元线性回归模型，即：

$$Y=\beta_0+\beta_1X+\varepsilon$$

当 $p>1$ 时，即影响被解释变量的主要因素在一个以上，模型中引入超过一个解释变量时，该模型被称为多元回归模型。如 $p=2$ 时的二元线性回归模型：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$$

（三）估计的回归方程

理论回归模型中的参数是未知的，回归分析的主要任务之一就是通过样本观测值（$y_i,x_{1i},x_{2i},\cdots,x_{pi}$）对 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 进行估计，我们可以用 $b_0,b_1,\cdots,b_p$ 分别表示 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 的估计值，则称：

$$\hat{Y}=b_0+b_1X_1+\cdots+b_pX_p$$

为估计的线性经验回归方程，或估计的线性回归方程。

其中，$\hat{Y}=b_0+b_1X_1$ 为估计的一元线性回归方程；$\hat{Y}=b_0+b_1X_1+b_2X_2$ 为估计的二元线性回归方程。

二、模型参数估计

利用所获取的样本数据，依照不同的准则，采用不同的统计方法，对回归模型中的参数进行估计，可以得到不同的参数估计值，即 $b_0,b_1,\cdots,b_p$ 不是唯一的。为了通过样本数据得到回归模型 $Y=\beta_0+\beta_1X_1+\cdots\beta_pX_p+\varepsilon$ 中 $\beta_0,\beta_1,,\cdots,\beta_p$ 的理想估计值，通常采用普通最小二乘法。普通最小二乘法的基本思想如下。

对每一样本观测值（$y_i,x_{1i},x_{2i},\cdots,x_{pi}$），考虑观测值 y_i 与其回归值（估计值）$\hat{y}_i$ 的离差越小越好，综合地考虑 n 个离差值，定义离差平方和为：

$$Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-(b_0+b_1x_{1i}+\cdots+b_px_{pi}))^2$$

所谓最小二乘法，就是要寻找 $\beta_0,\beta_1,\cdots,\beta_p$ 的估计值 $b_0,b_1,\cdots,b_p$，使 Q 达到最小。

求解 $b_0,b_1,\cdots,b_p$ 是一个求极值问题，由于 Q 是关于 $b_0,b_1,\cdots,b_p$ 的非负二次函数，因而它的最小值总是存在的。根据微积分求极值的原理，$b_0,b_1,\cdots,b_p$ 应满足下列方程：

$$\begin{cases}-2\sum_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_px_{pi})=0\\-2\sum_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_px_{pi})x_{1i}=0\\\vdots\\-2\sum_{i=1}^{n}(y_i-b_0-b_1x_{1i}-b_2x_{2i}-\cdots-b_px_{pi})x_{pi}=0\end{cases}$$

求解该方程组，即可得到$b_0,b_1,\cdots,b_p$。

对于一元线性回归方程，其参数估计值的具体计算公式为：

$$\begin{cases} b_1=\dfrac{n\sum_{i=1}^{n}xy-\sum_{i=1}^{n}x\sum_{i=1}^{n}y}{n\sum_{i=1}^{n}x^2-\left(\sum_{i=1}^{n}x\right)^2} \\ b_0=\bar{y}-b_1\bar{x} \end{cases}$$

回归模型参数估计的 SPSS 实现：

(1) SPSS 选项：【Analyze】—【Regression】—【Linear】。

(2) 在【Linear Regression】框中，输入消费支出到【Dependent】对话框，输入支配收入到【Independent】对话框中，单击【OK】按钮即可。输入结果如表 12-7 所示。

表 12-7 居民人均可支配收入与人均消费支出的一元线性回归（a）

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Eror	Beta		
1	(Constant)	450.334	388.906		1.158	.256
	人均可支配收入	.692	.029	.976	24.148	.000

a. Dependent Variable：人均消费支出。

表 12-7 给出了回归方程中常数项和回归系数的估计值。可以看到$b_0=450.334$，$b_1=0.692$，其回归的直线如图 12-3 所示。

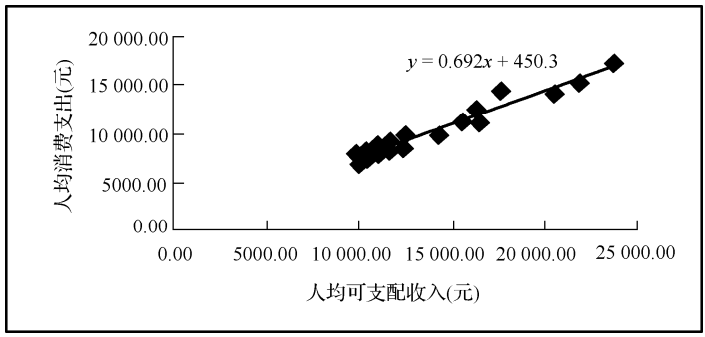


图 12-3 居民人均可支配收入与人均消费支出散点图与回归线

回归方程为：

$$Y(\text{年人均消费支出})=450.334+0.692X(\text{年人均可支配收入})$$

三、回归系数的含义

(一) 一元线性回归模型

$\hat{Y}=b_0+b_1X$ 中的 b_0 是直线的截距，表示当解释变量为零时 Y 的平均值。回归系数 b_1 是直线的斜率，表示解释变量 X 每增加一个单位，被解释变量将相应地平均变化 b_1 个单位。

上面回归方程中回归系数 b_1=0.692 的含义为：年人均可支配收入每增加 1 元，年人均消费支出会平均增加 0.692 元。

（二）多元线性回归模型

$\hat{Y}=b_0+b_1X_1+\cdots+b_pX_p$ 中的回归的 b_i (i=1,2,⋯,p) 称为偏回归系数，表示当模型中其他解释变量保持不变时，该解释变量增加一个单位，被解释变量相应地平均变化 b_i 个单位。

四、回归方程的评价与检验

当我们得到一个实际问题的经验回归方程后，还不能马上就得出分析结论或进行预测等实际应用。在应用前需要对所估计的回归方程进行评价与检验。进行评价与检验主要是基于以下理由：第一，在利用样本数据估计回归模型时，首先是假设变量 Y 与 $X_1,X_2,\cdots,X_p$ 之间存在着线性关系，X_i 对 Y 的关系是显著的，但这种假设是否存在需要进行检验；第二，用样本数据估计的回归方程是否真正描述了变量 Y 与 X_i 之间的统计规律性，Y 的变化能否通过模型中的解释变量 $X_1,X_2,\cdots,X_p$ 去解释需要进行评价等。一般进行的评价与统计检验的主要内容与方法有以下几点。

（一）实际意义检验

所谓实际意义的检验就是利用相关领域的基本常识、相关专业的原理及我们所积累的丰富的经验，对所估计的回归方程的回归系数进行分析与判断，看其是否能得到合理的解释。如我们以人均消费支出为被解释变量，以人均可支配收入为解释变量，估计的一元线性回归方程如下：

$$Y=450.33+0.692X$$

回归系数 0.692 的含义是人均可支配收入每增长 1 元，则人均消费支出将平均增加 0.692 元。其经济意义合理与经济理论、实际情况相符，但如果回归系数是负数，则与实际或经济理论不相符。

对回归模型进行检验，首先要进行的就是实际意义的检验。

（二）回归方程的拟合程度分析

回归方程在一定程度上描述了变量 Y 与 $X_1,X_2,\cdots,X_p$ 之间的数量依存关系与内在规律，根据这一方程，我们可由解释变量的取值来估计被解释变量的取值，但估计的精度如何将取决于回归方程对观测数据的拟合程度。回归方程的拟合程度的分析最常用的指标是判定系数。

1. *判定系数* R^2

判定系数是说明回归方程拟合程度的一个度量值，以一元线性回归方程为例，若各观测数据 (x_i,y_i) 在坐标系上形成的散点都落在一条直线上，那么这条直线就是对数据的完全拟合，直线充分代表了各个点所显示的变量之间的关系，此时，用 X 估计 Y 是没有误差的。各样本观测点越是紧密围绕直线，说明直线对观测数据的拟合程度越好，判定系数越高；反之，则越差，判定系数就会越小。

为理解判定系数的含义，我们可以通过对被解释变量取值的变差进行分析。

我们的研究对象被解释变量 Y 是个变量，其取值是不同的、有波动的，Y 取值的这种波动我们可以称之为变差。而变差的产生来自于两个方面：一是由于解释变量 X 的取值不同

造成的；二是除 X 外的其他因素的影响，包括随机因素的影响。对一个具体的观测值来说，变差的大小可以通过该实际观测值与其均值之差 $(y_i-\bar{y})$ 来表示。而 n 个观测值的总变差可以由这些离差的平方和来表示，称为总变差平方和，用 SST 表示，即 $\mathrm{SST}=\sum(y_i-\bar{y})^2$ 。

以一元线性回归方程为例，估计的回归方程为直线方程：$\hat{Y}=b_0+b_1X$ 。从图 12-4 中可以看到，每个观测点的离差都可以分解为两部分，即：

$$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$$

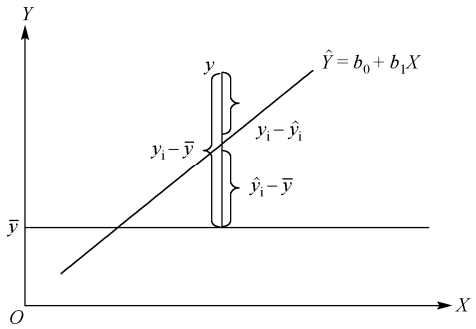


图 12-4　变差分解图

各观测数据独立时，将上式两边平方，并对所有 n 个点求和可以得到(证明略)：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

即总的变差平方和可以分解为两个部分：一部分是 $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ ，它是回归值 $\hat{y}_i$ 与 Y 的均值 $\bar{y}$ 的离差平方和，可以看作是 Y 的总变差中由于 X 与 Y 的线性关系引起的 Y 的变化的那部分，可以由回归直线来解释，因而称为可解释的变差平方和或回归平方和，记为 SSR；另一部分是 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ ，即 Y 的各实际观测点与其利用回归方程估计值的残差的平方和，它是除了 X 对 Y 的线性影响之外的其他因素形成 Y 的变差，是不能用回归直线来解释的，因而称为不可解释的变差或剩余平方和，记为 SSE。三个平方和的关系是：

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}$$

在观测值已知的情况下，利用 SPSS 软件进行回归分析的计算，其离差平方和的计算结构如表 12-8 所示。

表 12-8　离差平方和计算结果表

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	182436901	1	182436901.4	583.144	.000[a]
	Residual	9072666.8	29	312850.578		
	Total	191509568	30			

a. Predicators:（Constant），人均可支配收入。

b. Dependent Variable：人均消费支出。

上面 SPSS 输出结果显示出，SSR=182436901，SSE=9072666.8，SST=191509568。

从变差分解图中可以直观地看到，回归直线拟合的好坏取决于 SSR 及 SSE 的大小，各观察值越是靠近直线，SSR 就越大，即 SSR 占 SST 的比例就越大。这样我们可以通过这一比例来反映直线对观测值的拟合程度，这一比例被称为判定系数，记为 R^2，即：

$$R^2=\frac{\mathrm{SSR}}{\mathrm{SST}}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

判定系数 R^2 的取值范围在[0，1]，R^2=1 时，拟合是完全的，即所有观测值都在直线上，若 X 与 Y 无关，X 完全无助于解释 Y 的变差，此时 $\hat{y}_i=\bar{y}$，则 R^2=0。可见，R^2 越接近于 1，表明回归平方和占总变差平方和的比重越大，回归直线与各观测点越接近，用 X 的变化来解释 Y 的变差部分越多，回归直线的拟合程度就越好。反之，R^2 越接近于 0，回归直线的拟合程度越差。

下面例题中，人均消费支出与人均可支配收入间的一元线性回归方程的判定系数为 0.953（见表 12-9 中的判定系数计算结果）。其实际意义是：在人均消费支出的总变差中，有 95.3%可以由人均可支配收入与人均消费支出之间的关系来解释，可见回归方程的拟合程度较高。

表 12-9　判定系数计算结果

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.976	0.953	0.951	559.33047

2. 调整后的判定系数 $\overline{R}^2$

在多元线性回归分析中，同样需要计算判定系数进行拟合程度的评价。但是模型中的解释变量越多，对 Y 的变差的解释程度也就越高，则用上面公式计算的 R^2 也就越大。但是多引入一个解释变量在进行统计检验时就会减少一个自由度，从而导致参数估计的有效性降低，甚至可能由于解释变量之间的高度线性相关，导致出现参数估计结果与实际情况不相符的结果。因此在多元线性回归模型的估计中一般需要考虑到解释变量个数对 R^2 的影响，即对判定系数 R^2 进行处理，计算调整后的 $\overline{R}^2$（计算略）。对于 R^2 我们可以直接利用统计软件得到其计算结果，表 12-9 中，调后的判定系数（Adjusted R Square）$\overline{R}^2$=0.951。需要注意的是在应用时，若回归方程的常数项为 0 时，R^2 失效。

3. 估计标准误差

估计标准误差是回归方程残差的均方根，用公式表示为：

$$s_e=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-p-1}}=\sqrt{\frac{\mathrm{SSE}}{n-p-1}}$$

式中，n 表示样本量；p 表示回归方程中解释变量的个数。

s_e 反映了实际观测值 y_i 与回归估计值 $\hat{y}_i$ 之间的差异程度。从实际意义上看，s_e 反映了用估计的回归方程估计被解释变量时估计误差的大小。s_e 越小，实际观测值与估计值的差异越小，回归方程对各观测点的代表性就越好，拟合程度越高，根据各回归方程进行预测也就越准确。

（三）回归系数的显著性检验——*t* 检验

t 检验是统计推断中常用的一种检验方法，在回归分析中，*t* 检验主要用于检验回归系数的显著性，即解释变量对被解释变量的影响是否显著。

检验的假设是：

$$H_0:\beta_i=0\text{（解释变量对被解释变量的影响不显著）}$$

$$H_1:\beta_i\neq 0\text{（解释变量对被解释变量的影响显著）}$$

如果 H_0 成立，则可认为因变量 Y 与解释变量 X_i 之间并没有真正的线性关系，即 X_i 的变化对 Y 并没有显著的线性影响。

构造的检验统计量为：

$$t=\frac{b_i}{\sqrt{\mathrm{var}(b_i)}}$$

其中，$\sqrt{\mathrm{var}(b_i)}$ 为回归系数估计量 b_i 的标准差。

在显著性水平为 α 的条件下，若 $|t|\geq t_{\alpha/2}$ 时，拒绝原假设，认为 $\beta_i\neq 0$，即解释变量 X_i 对 Y 的线性影响显著；否则，认为 X_i 对 Y 的线性影响不显著。

若伴随概率（*P* 值或 Sig 值）小于我们事先确定的显著性水平 α 时，拒绝原假设，认为 $\beta_i\neq 0$，即解释变量 X_i 对 Y 的线性效果显著。

表 12-10 是居民人均可支配收入和人均消费支出回归分析的 SPSS 计算结果输出表。可以看出，解释变量人均可支配收入的回归系数的 *t* 检验统计量为 24.148，*P* 值为 0.00 小于 0.05，所以拒绝原假设，可以认为人均可支配收入与人均消费支出的线性影响显著。

表 12-10　回归系数及其统计检验结果

Model		Unstandardized Coefficients		Standardized Coefficients	*t*	Sig.
		B	Std. Eror	Beta		
1	(Constant)	450.334	388.906		1.158	.256
	人均可支配收入	.692	.029	.976	24.148	.000

a. Dependent Variable：人均消费支出

（四）回归方程线性关系的显著性检验——*F* 检验

回归方程线性关系的检验称为 *F* 检验，它用于检验解释变量 X_i 和被解释变量之间的线性关系式是否显著；或者说，它们之间能否用一个线性模型 $Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\varepsilon$ 来表示。*F* 检验是根据 *Y* 的离差平方和分解式，直接根据回归效果检验回归方程的显著性。如果是显著的，说明回归方程线性关系是存在的，如果不显著，说明回归方程的线性关系是不存在的。

F 检验的具体步骤如下。

首先，提出假设：

$$H_0:\beta_1=\beta_2=\cdots=\beta_p=0\text{（回归方程的线性关系不显著）}$$

然后，计算检验统计量：

$$F=\frac{\mathrm{SSR}/p}{\mathrm{SSE}/(n-p-1)}$$

可以证明，在原假设成立的情况下，F 检验统计量服从 F 分布，第一自由度为 p（解释变量的个数），第二自由度为 $n-p-1$（n 为观测数据个数），即 $F \sim F(p,n-p-1)$。

在给定的显著性水平 α 时，当检验统计量的数值大于 $F_\alpha(p,n-p-1)$ 时，拒绝原假设，说明回归方程的线性关系是存在的。

当伴随概率（P 值或 Sig 值）小于我们事前确定的显著性水平 α 时，拒绝原假设，认为 $\beta_1,\beta_2,\cdots,\beta_p$ 中至少有一个是不为零的，回归方程的线性关系是存在的。

在一元线性回归分析时，由于只有一个解释变量，因此 t 检验与 F 检验的结果是一致的。

表 12-11 是居民人均可支配收入与人均消费支出回归方程的方差分解和 F 检验计算结果。

表 12-11　变差分解及 F 检验计算结果

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	182436901	1	182436901.4	583.144	.000[a]
	Residual	9072666.8	29	312850.578		
	Total	191509568	30			

a. Predicators: (Constant)，人均可支配收入

b. Dependent Variable：人均消费支出

上表显示出居民人均可支配收入和人均消费支出回归方程的线性关系检验统计量 F 值为 583.144，P(Sig)值为 0.00 小于显著性水平 0.05，所以拒绝原假设，认为由解释变量和被解释变量建立的回归方程的线性关系是存在的。

五、利用回归方程进行预测

回归分析的主要目的是根据所建立的回归方程，用给定的解释变量来预测被解释变量，如果对于 $X_1,X_2,\cdots,X_p$ 的给定值，求出 Y 的一个预测值 $\hat{Y}$，就是点估计。在点估计的基础上，可以得到 Y 的估计区间。

估计区间有两种类型，即平均值的置信区间和个别值的预测区间。

（一）Y 的平均值的置信区间

平均值的置信区间是对于 $X_1,X_2,\cdots,X_p$ 的给定值，求出 Y 的平均值的估计区间。如本节例中根据人均可支配收入与人均消费支出的回归方程，估计当人均可支配收入为 25 000 元时人均消费支出平均值的估计区间。由于计算公式较为复杂，本节中只给出利用一元线性回归方程计算平均值的置信区间。

在 $X=x_0$，$1-\alpha$ 的置信度下，y_0 的平均值的置信区间计算公式如下：

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

式中，$s_e \sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$ 为 $\hat{y}_0$ 的标准差的估计量。

（二）Y 的个别值的预测区间

个别值的预测区间是对 $X_1,X_2,\cdots,X_p$ 的给定值，求出 Y 的一个个别值的估计区间，如当

某家庭人均可支配收入为 25 000 元时，根据居民人均可支配收入与人均消费支出的回归方程，估计该家庭人均消费支出的区间。

在 $X=x_0$，$1-\alpha$ 的置信度下，y_0 的个别值的置信区间计算公式如下：

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

可以看到，即使是解释变量的值及置信水平相同，这两个区间的宽度也是不一样的，预测区间要比置信区间宽一些。

用 SPSS 进行回归预测计算：

(1) SPSS 选项：【Analyze】—【Regression】—【Linear】。

(2) 在【Linear Regression】框中，输入消费支出到【Dependent】对话框，输入支配收入到【Independent】对话框中。

(3) 单击【save】按钮。在【predicted values】下选中【Unstandardized】，在【predection Interval】下选【Mean】和【Individual】即可。

本例中的预测值(表 12-12)及图形(图 12-5)如下。

表 12-12 人均消费支出预测值

人均可支配收入	人均消费支出	地区生产总值	LMCI_1	UMCI_1	LICI_1	UICI_1
10012.34	7875.78	2702.40	7104.35915	7652.81599	6202.21857	8554.95657
10245.28	7519.28	7065.00	7274.39140	7805.15940	6365.43667	8714.11414
10276.06	7512.39	783.61	7296.82834	7825.32023	6386.99221	8735.15636
10313.44	7874.27	3523.16	7324.06630	7849.81405	6413.16615	8760.71420
10678.40	7758.69	2741.90	7589.39089	8089.57327	6668.50454	9010.45962
10763.34	8427.06	5465.79	7650.97130	8145.54497	6727.87633	9068.63994
10859.33	7817.28	889.20	7720.47943	8208.88152	6794.94677	9134.41418
10996.87	8292.89	1223.28	7819.91390	8299.79455	6891.00259	9228.70587
11098.28	8691.99	10505.30	7893.10161	8366.95250	6961.79058	9298.26353
11130.93	7532.07	342.19	7916.64149	8388.50835	6984.57514	9320.66479

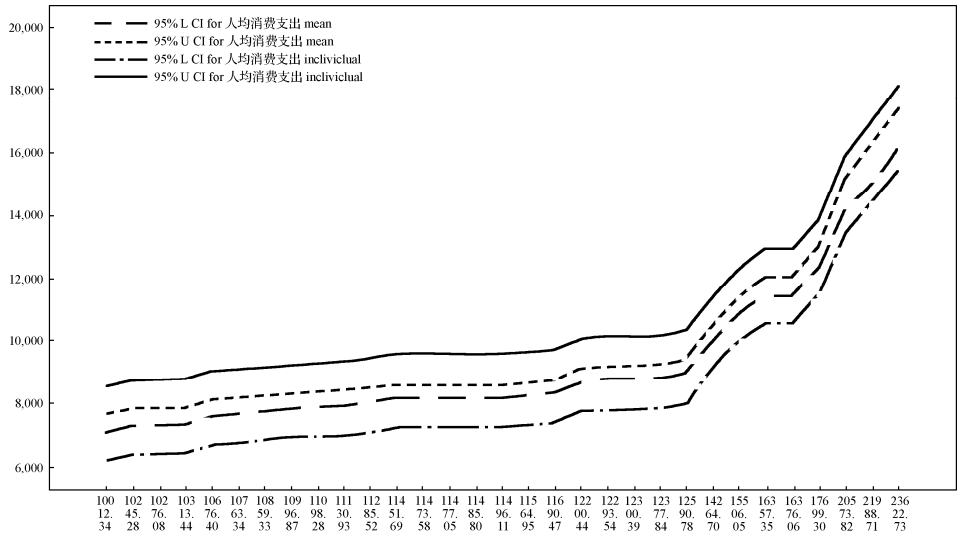


图 12-5 人均消费支出预测值

表 12-12 中 LMCI_1 和 UMCI_1 分别表示 Y 的平均值的置信区间的上、下限，而 LICI_1 和 UICI_1 分别为 Y 的个别值的预测区间的上、下限。

第三节　可线性化的非线性回归

一、可线性化的非线性回归模型

在许多实际问题中，变量之间的关系并不都是线性的。通常我们会碰到某些现象的被解释变量与解释变量之间呈现某种非线性关系。如对于两个变量 X 和 Y，若被解释变量 Y 随解释变量 X 的取值的不同而变化，并且呈现出某种曲线形态时，我们称两者之间存在非线性关系，这时应采用适当的曲线来描述两者之间的关系。

在只涉及一个解释变量时，称两个变量之间的回归分析为一元非线性回归，若涉及多个解释变量的回归分析，我们将其称为多元非线性回归分析。

在非线性回归分析时，有相当多的回归方程是可以通过简单的变量变换，使其转换为线性模型。这时，便可将非线性回归问题转化为线性回归问题并进行参数的估计、模型的拟合和分析，我们将这类回归称为可线性化的非线性回归，本节只介绍几种常见的可线性化的非线性回归模型。

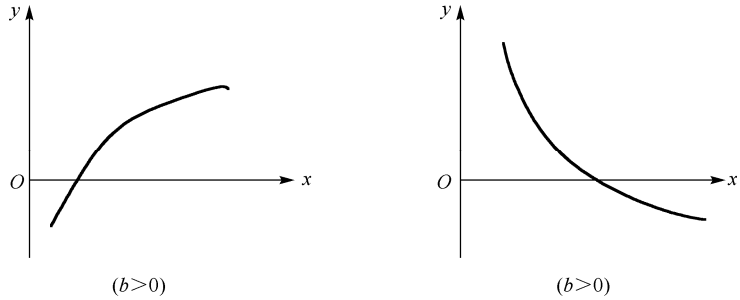
二、主要模型及参数估计

（一）对数曲线（Logarithm）

对数曲线方程：

$$Y = b_0 + b_1 \ln X$$

曲线的图形如图 12-6 所示。



($b>0$)　　($b>0$)

图 12-6　对数曲线

模型参数的估计方法如下。

首先，进行变量变换，将曲线模型转化为线性模型，对此模型线性化的方法是：

令 $X' = \ln X$，则有 $Y = b_0 + b_1 X'$。

然后，利用最小二乘法估计模型 $Y = b_0 + b_1 X'$ 的参数即可。

（二）逆函数曲线（inverse）

若变量 Y 随 X 的增加而增加，最初增加很快，以后逐渐减慢并趋于稳定，则可以选择逆函数曲线，其回归方程为：

$$Y = b_0 + b_1 \frac{1}{X}$$

曲线的图形如图 12-7 所示。

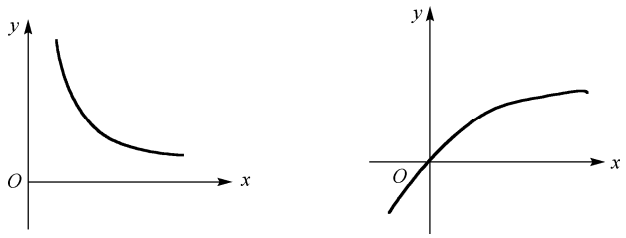


图 12-7　逆函数曲线

模型参数估计方法如下。

首先，进行变量变换，将曲线模型转化为线性模型，对此模型的线性化方法是：

令 $X' = \frac{1}{X}$，则有 $Y = b_0 + b_1 X'$。

然后利用最小二乘法估计模型参数即可。

（三）二次曲线（Quadratic）

二次曲线是非常常见的一种非线性回归形式，它的方程是：

$$Y = b_0 + b_1 X + b_2 X^2$$

如果令 $X_1 = X$，$X_2 = X^2$，则上面二次曲线方程线性化为二元一次方程：

$$Y = b_0 + b_1 X_1 + b_2 X_2$$

之后，采用最小二乘法便可估计模型中未知参数。

（四）三次曲线（Cubic）

三次曲线方程：$Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3$。

令 $X_1 = X, X_2 = X^2, X_3 = X^3$，则模型线性化为：

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3$$

（五）幂函数曲线（Power）

若变量 X 与 Y 都接近等比变化，即其环比分别接近于一个常数，可配合幂函数曲线，其方程为：$Y = aX^b$。

该模型的线性化方法如下。

将模型两端取对数，得到：$\ln Y = \ln a + b \ln X$。

然后，令 $Y' = \ln Y, X' = \ln X, A = \ln a$，则有：

$$Y' = A + bX'$$

利用最小二乘法估计参数 A 和 b 后，对 A 求反对数则可以得到模型中的参数 a。

例 12-1：某出租车公司随着出租车数量的增加，每年发生交通事故造成的损失也在增加，具体收集到的数据资料如表 12-13 所示。

表 12-13　出租车数量与年损失金额

年份	1	2	3	4	5	6	7	8	9	10
出租车数量(百辆)	40	42	48	55	65	79	88	100	120	140
年损失金额(百元)	3000	2800	32 000	34 000	3000	3240	3700	3300	3800	3700

SPSS 实现过程：

(1)建立数据文件。定义变量，y 表示年损失金额，x 表示出租车数量。

(2)在数据文件管理窗口中选择【Graphs】，展开下拉菜单单击 Scatter 按钮，进入【Scatter plot】对话框，单击【Simple】按钮，再单击【Define】按钮，在【Define】对话框中，将变量 y 放入【Y Axis】栏，变量 x 放入【X Axis】栏，单击【OK】按钮。得到散点图(如图 12-8 所示)，从散点图中可看出两变量呈现某种曲线关系。

(3)在数据文件管理窗口中选择【Analyze】，展开下拉菜单，单击【Regression】中的【Curve Estimation】，进入【Curve Estimation】对话框，把左侧变量栏的变量 y 放入【Dependents】栏，变量 x 放入【Independent】栏中。从 Models 栏中选择【Quadratic】、【Cubic】和【Power】三个复选项，如图 12-9 所示。

(4)选择完后，单击【OK】按钮，得到计算结果，如表 12-14 所示。

相关的输出结果如下。

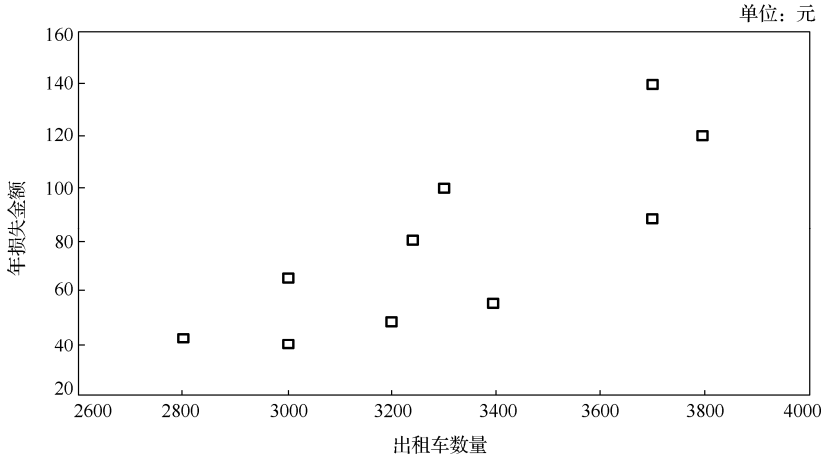


图 12-8　年损失额与出租车数量之间的散点图

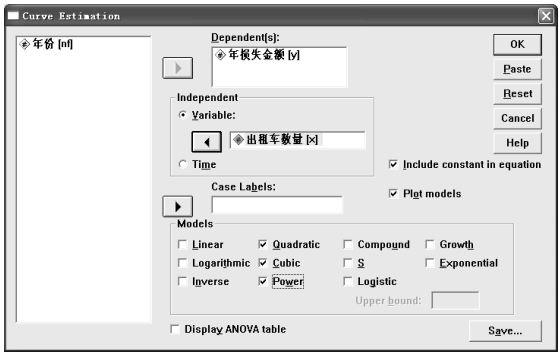


图 12-9　曲线估计对话框

表 12-14　三种曲线的拟合结果

	Model Summary					Parameter Estimates			
Equation	R Square	*F*	df1	df2	Sig.	Constant	b1	b2	b3
Quadratic	.148	.607	2	7	.571	24747.649	-266.151	.731	
Cubic	.270	.741	3	6	.565	-72431.2	3704.309	-47.937	.182
Power	.093	.817	1	8	.392	93275.306	-.675		

Dependent Variable: 年损失金额

The independent variable is 出租车数量.

第四节　相关与回归分析应用中的几个问题

一、建立回归模型的基本过程

（一）根据研究目的，设置指标变量

回归分析模型主要是反映被解释变量与相关的解释变量之间的数量依存关系。为此，首先应根据所研究问题的目的来确定被解释变量和解释变量。

被解释变量的确定，基本要求是其应能够表达、刻画研究对象的基本特征。一般当研究对象、研究目的确定后，被解释变量能够较为容易的确定。如要研究地区通货膨胀变动的规律，就可以考虑将零售物价指数作为被解释变量。

而解释变量的确定，相对于被解释变量难度要大一些，这也是回归分析应用中的难点与重点。确定解释变量，常常需要对所研究的对象、被解释变量的变化特征及其背景要有足够的了解，如以股票价格指数为被解释变量，则影响股票价格指数的有关因素就可以考虑作为解释变量。

在确定解释变量时，应注意到并不是模型所涉及的解释变量越多越好。建立回归模型时，如果漏掉一些重要的解释变量肯定会影响模型的效果，但如果引入过多的解释变量，也会出现很严重的问题。首先是计算量加大，但更为严重的是可能会出现解释变量之间的信息相互影响，参数估计的有效性降低，甚至会出现回归系数的估计值的符号与实际状况相反的情况。

（二）收集、整理统计数据

回归模型的建立是基于回归变量的样本统计数据，在确定了被解释变量与解释变量后，就要收集、整理相关的统计数据，数据是建立回归模型的重要的、基础性的工作，数据资料的质量，会直接影响到所估计的回归模型的效果与应用。

建立回归模型时最基础的、常用的样本数据分为时间数列数据和横截面数据，其中时间数列数据是将被解释变量及解释变量的数据按其发生的时间先后顺序排列的统计数据，使用时间数列数据时应特别注意数据的可比性；而横截面数据是在同一时间截面上的统计数据。建立回归模型时，不论是时间数列数据还是截面数据，其样本量都要足够的大。英国统计学家 M·肯德尔在《多元分析》中指出，样本容量 n 应是解释变量个数的 10 倍，但在实际应用时，可能达不到这一数量。统计数据不全是普遍的事，我们在收集数据时要尽可能地多收集一些样本数据，并对样本数据中的异常值进行处理。

（三）确定理论回归模型的数学形式

若解释变量只有一个时，确定理论回归模型的数学形式，可以直接利用样本的散点图来

选择，若样本点大致分布在一条直线的周围，我们可考虑用线性回归模型去拟合这条直线，即选择一元线性回归模型。

若解释变量有多个，通常要依据经济理论和一些数理经济学模型来确定理论回归模型的数学形式，如要建立一个关于产出的回归模型，如工业生产和农业生产的产出问题时，可以依据生产函数 $Y = AK^{\alpha}L^{\beta}$，其中 K 表示资本的投入，L 表示劳动力的投入。

有时，我们无法依据所获取的信息确定模型的形式，这时可以采用不同的形式进行模拟，对不同的模拟结果，选择较好的一个作为理论模型。

(四)模型的参数估计与检验

利用所收集的样本数据对模型的未知参数进行估计是回归分析的一项重要内容，未知参数的估计在不同的条件下有不同的方法。在众多的方法中，普通最小二乘法是最基本的、最常用的方法，但是对于不满足模型基本假设的回归问题，用最小二乘法估计出的参数，其估计的效果会受到影响。

未知参数估计后，还需要对其进行检验与评价来确定，能否使用该模型进行经济问题的分析与研究，以判定这个模型是否真正揭示了被解释变量与解释变量之间的关系，最常进行的检验与评价包括：

(1)实际意义的检验；

(2)拟合程度分析，如 R^2 和估计标准误差的计算;

(3)统计检验，如 t 检验、 F 检验等。

(五)模型的应用

回归模型通过了各种统计检验，且模型具有合理的实际意义时，就可以将其应用于实际分析中了。回归模型的主要应用有：

(1)利用模型的回归系数发现经济变量之间的结构关系，给出政策评价的一些量化依据；

(2)回归模型揭示了变量之间的因果关系，在给定被解释变量值时，可以用来控制解释变量的数值；

(3)利用回归模型进行预测分析。

二、解释变量的确定与筛选方法

建立回归模型时重要的问题之一是解释变量的选择，在确定解释变量时，要依据两条重要的准则：一是选择的解释变量应是与被解释变量之间密切相关的因素；二是所选择的解释变量之间不能有较强的线性关系。一般，解释变量选择的方法主要有以下几个。

(一)定性分析法

定性分析法是凭借对研究对象的熟悉、了解，分析找到影响被解释变量变化的主要因素，再从中选择那些能够定量描述且可搜集到观察值的因素，作为初选的解释变量。利用定性分析选择的解释变量要尽可能做到不遗漏重要的解释变量，将影响被解释变量的主要因素考虑全面。如考虑影响城镇居民消费支出额的主要影响因素，我们可以根据我们所熟悉的消费者消费行为、经济学中的消费理论等，分析出主要的影响因素包括消费者的可支配收入、家庭人口数、家庭资产、预期收入水平、过去时期的消费水平、历史上的收入水平、地区平均收入、零售价格指数、学历、年龄、家庭结构等。至于在模型中应引入哪些变量，需要进行进一步的筛选。

（二）相关分析法

众多的影响因素能否进入回归模型，还需要依据选择变量的基本准则，利用定量的方法进行筛选，筛选的方法之一是进行相关分析。

进行相关分析时，应包括以下两部分内容。

首先，分别计算被解释变量与各影响因素的简单相关系数，选择那些与被解释变量相关程度较高的变量作为解释变量。

其次，所选择的与被解释变量相关程度高的解释变量是否能全部进入回归模型，还取决于解释变量之间是否有较强的线性关系。因为在多元线性回归方程参数估计时，用最小二乘法得到的参数估计值 b_i 表示在其他解释变量保持不变时，由于 X_i 变动引起的 Y 的平均变动量。若解释变量间有较密切的线性关系，变量 X_i 稍有变化，与其高度相关的解释变量会随之变化，回归 b_i 将无法真正解释各解释变量对被解释变量的影响。因此，当解释变量之间存在高度线性相关关系时，只能保留其中一个，通常保留与被解释变量相关程度高的变量。

（三）逐个剔除法

逐个剔除法是先将与被解释变量有关的全部变量都当作解释变量引入到方程中，建立模型，然后利用回归系数的检验，依据每个回归系数的 t 检验统计量值的大小，逐个剔除那些不显著的变量，直到模型中包含的解释变量都是影响被解释变量的显著的因素为止。当不显著的解释变量多于一个时，不能将它们剔除掉，而是要首先剔除掉 t 的绝对值最小的那个变量，若删除这一变量后，模型的拟合效果更好，则认为该变量从模型中删掉是正确的。之后再利用剔除掉这一变量之后的变量再建立模型进行检验、剔除变量，一直到模型的回归系数检验通过为止。

（四）逐步回归法

逐个剔除法是一种常用的解释变量的筛选方法，但其有时会存在一定的问题，这主要是由于剔除掉的变量一般不会再引入到模型中。然而有时也许当后面剔除掉另一变量时，先前剔除掉的变量的回归系数的显著性检验可能就会通过，而证明该变量不应剔除。为了避免这一问题的出现，可以采用逐步回归分析筛选解释变量。

逐步回归分析的基本思想是，将各影响因素变量逐个引入回归模型，引入的条件是模型能够通过相关的统计检验，引入每个变量后，要对已引入的变量进行逐个检验，再剔除检验通不过的变量，一直到所有变量的回归系数均逐步做过分析之后。逐步回归分析的计算工作量较大，一般可以直接用统计软件实现。

三、带有定性解释变量的回归模型

在实际问题的研究中，影响被解释变量的因素中除了可以表现为数值型变量的影响因素外，还经常会碰到一些影响因素表现为非数值型的解释变量，如性别、学历、行业等。在建立一个回归方程时，经常需要考虑将这些特殊的变量引入到模型中作为解释变量。引入时，首先对这些非数型变量的情形进行数量化处理，处理的方法将其转化为只取 0 或 1 两个值的变量，即在某一属性出现时，该变量取值为 1，否则取值为 0，我们将其称为虚拟变量或哑变量。

(一)非数量型变量只取两类可能的值

有些非数值型变量只取两类可能的值，如若建立关于生产效率的回归模型，选择的解释变量中包括年龄 X_1、受教育年限 X_2 和性别，其中性别是虚拟变量，我们用 D 表示，可以将其数量化为：

$$D=\begin{cases}1 & \text{性别为男}\\ 0 & \text{性别为女}\end{cases}$$

这样，估计的回归模型就可表示为：

$$\hat{Y}=b_0+b_1X_1+b_2X_2+b_3D$$

b_0,b_1,b_2,b_3 值的计算可以采用最小乘法，变量 D 可以直接看作是一个取值为 1 或 0 的普通的变量即可。

(二)非数量型变量取值有多类

有些非数值型变量的取值可能不只两类，建立消费支出模型，选择的影响消费支出的变量包括收入 X_1、家庭资产 X_2、年龄 X_3 和学历，解释变量中学历水平分为低学历、中等学历和高学历等情况，即该变量取值有三类。这时在模型中引入两个虚拟变量，即：

$$D_1=\begin{cases}1 & \text{高学历}\\ 0 & \text{非高学历}\end{cases}$$

$$D_2=\begin{cases}1 & \text{中等学历}\\ 0 & \text{非中等学历}\end{cases}$$

这样估计的回归模型为：

$$\hat{Y}=b_0+b_1X_1+b_2X_2+b_3D_1+b_4D_2$$

当 $D_1=1$ 时，模型表示高学历者消费支出；

当 $D_2=1$ 时，模型表示中等学历者消费支出；

当 $D_1=0$，$D_2=0$ 时，模型表示低学历者消费支出。

若在模型中引入三个虚拟变量，则模型中不包括常数项。

四、回归分析应用——交通事故状况与机动车情况相关分析

(一)背景介绍

经过 20 多年来持续、稳定、高速的经济发展，中国的经济规模和综合国力已经有了极大的提高。作为国民经济的基础产业，交通运输基础设施建设和整个交通运输业成为经济生活中最为活跃的领域之一，并带动了社会生活的各个方面迅速发展，其中，全国城市道路及公路总里程、注册机动车辆数和驾驶员人数随之大幅、持续增加，道路交通流量不断上升，交通拥挤和交通事故频发。在经济迅猛发展的大背景下，机动车总量的增加无疑成为交通事故频发的重要因素之一。那么，能否通过运用某种统计分析方法来探寻交通事故发生的状况与机动车辆状况之间存在的关系呢？答案是肯定的，本例通过实际操作对简单相关分析及偏相关分析做了演示。

本例数据选中表示交通事故状况的变量可以有交通事故起数、死亡人数、受伤人数和损失折款；表示机动车情况的相关变量有民用汽车总量、新注册载客汽车量、私人载客汽车量和公路运输汽车拥有量。

(二)统计分析过程及技术实现

从变量所代表的意义上看，我们知道交通事故的损失与发生交通事故的数量密切相关，因此我们首先绘制散点图来观察他们之间的关系。

在 SPSS 中选择【Graphs】—【Legacy Dialogs】—【Scatter/Dot】—【Simple Scatter】，单击 Define 按钮打开简单散点图对话框:

(1)将交通事故损失折款选入 *Y* 轴;

(2)将交通事故发生起数选入 *X* 轴。

单击 OK 按钮即可得到散点图(如图 12-10 所示)。

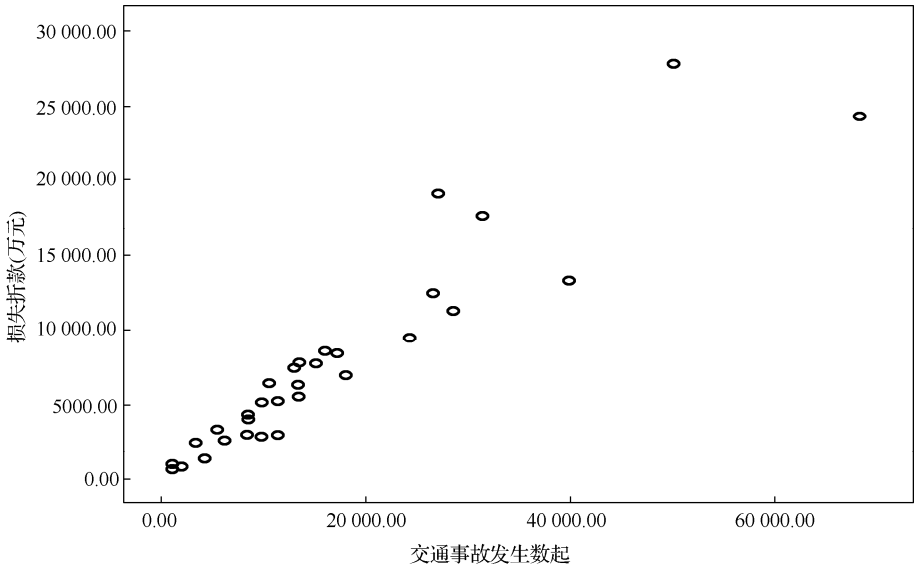


图 12-10 散点图

从图 12-10 中不难发现，损失折款与交通事故发生数呈现较为明显的正相关关系，那么相关程度到底有多大呢？必须通过相关系数才能得到，相关系数比相关图更能概括说明相关的形式和程度，为此我们还需要进行相关系数的计算并通过统计检验才能确定这种关系。利用 SPSS 计算简单相关系数的操作步骤如下。

在 SPSS 中选择【Analyze】—【Correlate】—【Bivariate】。

(1)将损失折款与交通事故发生数选入【variables】;

(2)在【correlation coefficient】中系统默认为 Pearson 相关系数;

(3)采取系统默认的双尾检验;

(4)单击【OK】按钮即可得到相关系数表(如表 12-15 所示)。

得到相关系数矩阵，如表 12-15 所示。

相关系数 r=0.933，显著性水平 0.000，这表明交通事故的损失折款与发生次数高度正相关。

此外我们还知道，在交通事故损失折款中包括了财产物品损失以及由于人员伤亡造成的损失，为此我们希望知道在控制伤亡人数的条件下损失折款与事故发生数的关系，这样就需要计算偏相关系数。其操作步骤如下。

表 12-15　相关性

		交通事故发生起数	损失折款万元
交通事故发生起数	Pearson 相关性	1	.933**
	显著性（双侧）		.000
	N	31	31
损失折款万元	Pearson 相关性	.933**	1
	显著性（双侧）	.000	
	N	31	31

**.　在 .01 水平（双侧）上显著相关。

在 SPSS 中选择【Analyze】—【Correlate】—【Partial】。

(1)将变量损失折款与交通事故发生数选入【variables】；

(2)将变量受伤人数与死亡人数作为控制变量选入【Controlling】；

(3)其他各选项选取系统默认值;

(4)单击【OK】按钮。

于是得到在控制了受伤人数与死亡人数之后交通损失折款与事故发生数的相关系数如表 12-16 所示。

表 12-16　相关性

控制变量			损失折款万元	交通事故发生起数
死亡人数 人 & 受伤人数 人	损失折款万元	相关性	1.000	.850
		显著性（双侧）	.	.000
		df	0	27
	交通事故发生起数	相关性	.850	1.000
		显著性（双侧）	.000	.
		df	27	0

从表 12-16 中可以看到，在控制了死亡人数和受伤人数之后损失折款与交通事故发生数间的相关系数发生了变化，由原来的 0.933 变为了 0.850，依然是高度正相关的。这与实际情况非常吻合，交通事故越多其损失必然越大。但在有些情况下，控制其他变量后所计算得到的偏相关系数并不一定还如简单相关系数那样显著，即由于存在伪相关，当控制其他变量后相关程度便不再显著。

五、回归分析应用——交通事故损失影响因素的回归分析

(一)背景介绍

在上述交通事故发生情况的基础上，我们希望找寻交通事故损失的影响因素并构造其回归方程，故在上述数据之上保留原有的交通事故损失变量并将其作为因变量进行分析，而把新注册载客汽车数量作为影响交通事故损失的自变量之一保留；同时还需要分析各方面可能的其他影响因素。初步分析看来，各地区的交通损失与各地经济发展状况有一定联系，经济较为发达的地区发生交通事故的损失也必然比较惨重；此外，各地的道路面积、交通用地情况也会与交通事故损失情况有一定联系，道路面积的扩宽能否减少交通事故发生的损失呢？公路客运量的增加是否对交通事故的损事造成了影响？考虑到数据的可获得性以及经济水平因素、道路交通状况和公路客运情况的影响，最终引入了地区生产总值、公路客运总量、年末实有道路面积和交通用地等变量，拟采用回归分析方法研究交通事故的损失与这些变量间

的数量关系。

(二)统计分析过程及技术实现

在上述数据中，我们首先需要确定这些自变量与交通事故损失是否有关，故首先进行交通损失与各自变量的相关系数的计算。操作步骤如下。

在 SPSS 中选择【Analyze】—【Correlate】—【Bivariate】。

(1)将交通损失变量 y 及前述影响因素变量 $x1$-$x5$ 选入【variables】，单击【Paste】按钮将“VARIABLES=y v1 v2 v3 v4 v5”改为“VARIABLES=y with v1 v2 v3 v4 v5”；

(2)选取系统默认的 Pearson 相关系数及双尾检验;

(3)单击工具栏中【run current】键即可运行。

表 12-17 是计算得到的相关系数，从表中可以看到，在 0.05 的显著性水平下，新注册载客汽车数量、地区生产总值、年末实有道路面积、交通用地和公路客运量与交通损失都是显著正相关的。进一步分析不难发现，这些自变量间是存在相关性的，同时计算地区生产总值与道路交通面积、交通用地之间又存在较高的相关关系[计算得到的相关系数结果分别为 0.959 和 0.619(计算略)]。为避免伪回归问题的出现，我们可以排除其他变量的干扰进行偏相关分析。

表 12-17 相关性

		新注册载客汽车万辆	地区生产总值	年末实有道路面积	交通用地万公顷	公路客运量万人
交通事失故损失万元	Pearson 相关性	.801**	.849**	.771**	.368*	.714**
	显著性（双侧）	.000	.000	.000	.041	.000
	N	31	31	31	31	31

**. 在 .01 水平（双侧）上显著相关。

*. 在 0.05 水平（双侧）上显著相关。

在控制地区生产总值的条件下，计算因变量交通事故损失与实有道路面积、交通用地的偏相关系数操作如下。

在 SPSS 中选择【Analyze】—【Correlate】—【Partial】:

(1)将交通损失变量 y 及实有道路面积、交通用地变量 y，x3，x4 选入【variables】；

(2)将地区生产总值 x2 选入【controlling】；

(3)其他选择系统默认值;

(4)单击【OK】按钮即可。

得到如下所示的偏相关系数(表 12-18)。

表 12-18 相关性

控制变量			交通事失故损失(万元)	年末实有道路面积	交通用地（万公顷）
地区生产总值	交通事失故损失(万元)	相关性	1.000	-.289	-.377
		显著性（双侧）	.	.121	.040
		df	0	28	28
	年末实有道路面积	相关性	-.289	1.000	-.111
		显著性（双侧）	.121	.	.559
		df	28	0	28
	交通用地(万公顷)	相关性	-.377	-.111	1.000
		显著性（双侧）	.040	.559	.
		df	28	28	0

从表 12-18 中看到，在控制了地区生产总值变量后，年末实有道路面积与交通事故损失的相关系数为–0.289 且不再显著相关；交通用地与事故损失的相关系数为–0.377，在 0.05 的水平下是显著负相关的。从偏相关系数与简单相关系数的求解可以看到，其值差异是很大的，故在实践过程中需特别注意其他变量的影响作用。

为了构造交通事故损失的回归方程，我们将上述数据中的各自变量逐个引入其中，即采用逐步回归的方法建立线性回归方程，探寻各自变量对交通事故损失的影响作用是否显著。具体操作步骤如下：

在 SPSS 中选择【Analyze】—【Regression】—【Linear】：

(1)将交通事故损失 y 选入【Dependent】；

(2)将反映地区车辆状况、经济发展状况、道路交通状况、公路客运量的变量 x1-x5 作为自变量选入【Independts】；

(3)在【Method】中选择逐步回归法【Stepwise】；

(4)单击【OK】按钮。

运行后的输出结果如表 12-19、表 12-20 和表 12-21 所示。

表 12-19　模型摘要[d]

模型	R	R 方	调整的 R 方	估计的标准差	Durbin-Watson
1	.849[a]	.720	.711	3613.09085	
2	.872[b]	.760	.743	3405.19702	
3	.891[c]	.794	.771	3211.95971	2.383

a. 预测变量:(常量)，地区生产总值。

b. 预测变量:(常量)，地区生产总值，交通用地万公顷。

c. 预测变量:(常量)，地区生产总值，交通用地万公顷，公路客运量万人。

d. 因变量：交通事故损失万元。

表 12-20　ANOVA[d]

模型		平方和	df	均方	F	显著性
1	回归	974782769	1	974782769	74.671	.000[a]
	残差	378578339	29	13054425		
	合计	1.35E+009	30			
2	回归	1.03E+009	2	514345419	44.358	.000[b]
	残差	324670270	28	11595367		
	合计	1.35E+009	30			
3	回归	1.07E+009	3	358270203	34.727	.000[c]
	残差	278550500	27	10316685		
	合计	1.35E+009	30			

a. 预测变量:(常量)，地区生产总值。

b. 预测变量:(常量)，地区生产总值，交通用地万公顷。

c. 预测变量:(常量)，地区生产总值，交通用地万公顷，公路客运量万人。

d. 因变量：交通事故损失万元。

表 12-19 和表 12-20 分别表示模型拟合过程的摘要表和方差分析表。

模型摘要表显示经过三步得到最终的回归方程，其拟合优度 R^2=0.794，表明方程的拟合程度较好。

方差分析表为回归拟合过程中每一步的方差分析结果。表 12-20 中显示，F 统计量的值为 34.727，显著性水平 0.000，表明回归方程通过统计检验。

表 12-21 回归系数计算结果表

模型		非标准化系数		标准化系数			B 的 95% 置信区间	
		B	标准误差	Beta	t	显著性	下限	上限
1	（常量）	836.719	1026.919		.815	.422	-1263.565	2937.004
	地区生产总值	1.306	.151	.849	8.641	.000	.997	1.615
2	（常量）	2739.962	1309.902		2.092	.046	56.750	5423.175
	地区生产总值	1.548	.181	1.006	8.537	.000	1.177	1.919
	交通用地万公顷	-441.019	204.538	-.254	-2.156	.040	-859.996	-22.043
3	（常量）	2611.607	1237.058		2.111	.044	73.373	5149.841
	地区生产总值	1.280	.213	.832	6.009	.000	.843	1.717
	交通用地万公顷	-570.564	202.426	-.329	-2.819	.009	-985.908	-155.221
	公路客运量万人	.047	.022	.294	2.114	.044	.001	.093

a. 因变量：交通事故损失万元

从上面的回归系数表 12-21 中得到最终的方程将变量地区生产总值、交通用地和公路客运量选入，而排除了新注册载客汽车数量和年末实有道路面积两个指标。在系数表中，偏回归系数的显著性由 t 检验得到，在显著性水平 α=0.05 的条件下，地区生产总值、交通用地和公路客运量的偏回归系数均通过统计检验。除了回归系数的估计值外，我们还可以得到95%的置信区间。此外，标准化回归系数反映了数据经过标准化后对因变量的影响程度，可以看到地区生产总值对交通损失的影响最为明显。

（三）分析结果解读

通过对各地区交通事故的损失情况的相关与回归分析看到，在各项影响因素中，地区生产总值、交通用地和公路客运量对交通事故的损失有显著影响。经过回归分析得到的拟合方程为：

$$Y = 2611.607 + 1.286X_2 - 570.564X_4 + 0.047X_5$$

它表明：在其他变量保持不变的条件下，地区生产总值每增加 1 万元，发生交通事故时会引起交通事故损失平均增加 1.286 万元；交通用地每增加 1 万公顷，会使得交通事故的损失平均减少 570.564 万元；公路客运量每增加 1 万人，发生交通事故时就会引起平均 0.047 万元的交通损失。

(1) 地区生产总值对发生交通事故时的损失情况有显著影响。地区生产总值反映了各地区的经济发展状况，经济发达的地区人流、物流活动都比较频繁，一旦发生交通事故财务经济损失和人身事故损失非常严重。因此，当发生交通事故时，经济发达地区的各项损失往往比欠发达地区的损失要严重许多，其所造成的不良影响也比较深远。

(2) 交通用地的情况对交通事故损失的影响也是比较显著的。交通用地的增加会改善道路的交通状况，从而降低道路交通拥挤等造成事故多发的因素，进而减少了交通事故发生的概率，那么交通事故的损失自然得到了有效的控制。

(3) 公路客运量也是造成交通事故损失严重的一大重要影响因素。我们知道，许多特大交通事故发生在高速路客运过程中，发生交通事故会导致众多人员伤亡，从而引起较大的间接经济损失；此外，客运交通事故多发于投资成本较高的高速路上，对道路设施的破坏严重，这也会进一步加大发生交通事故的损失。

思考与练习

1．相关分析的主要目的和分析的内容有哪些？

2．回归分析的主要目的和分析的内容有哪些?

3．相关分析和回归分析的关系是什么？说明其主要区别。

4．什么是相关关系？它与函数关系有何不同？

5．从某行业中随机抽取 12 家企业，对其产量和生产费用进行调查，数据如下：

企业编号	产量(万台)	生产费用(万元)
n	*x*	*y*
1	40	130
2	42	150
3	50	155
4	55	140
5	65	150
6	78	154
7	84	165
8	100	170
9	116	167
10	125	180
11	130	175
12	140	185
合计	1025	1921

要求：

(1)根据数据绘制散点图，判断产量与生产费用之间的关系形态；

(2)计算产量与生产费用之间的相关系数，并说明二者之间关系的密切程度；

(3)对相关系数的显著性进行检验(α=0.05)。

6．搜集我国各省、市、自治区 2016 年的人均国内生产总值(GDP)和人均消费水平的统计数据，要求：

(1)绘制人均 GDP、人均消费水平的散点图，并说明二者之间的关系形态；

(2)计算两个变量之间的线性相关系数，说明两个变量之间的关系强度；

(3)利用最小二乘法求出估计的回归方程，并解释回归系数的实际意义；

(4)计算判定系数，并解释其意义；

(5)检验回归方程线性关系的显著性(α=0.05)；

(6)如果某地区的人均 GDP 为 50 000 元，预测其人均消费水平；

(7)求人均 GDP 为 50 000 元时，人均消费水平 95%的置信区间和预测区间。

7．随机抽取的 10 家航空公司，对其最近一年的航班正点率和顾客投诉次数进行了调查，所得数据见下表：

航空公司编号	航班正点率(%)	投诉次数(次)
n	*x*	*y*
1	81.8	21
2	76.6	58

续表

航空公司编号	航班正点率(%)	投诉次数(次)
3	76.6	85
4	75.7	68
5	73.8	74
6	72.2	93
7	71.2	72
8	70.8	122
9	91.4	18
10	68.5	125
合计	758.6	736

要求：

(1)绘制散点图，说明二者之间的关系形态；

(2)用航班正点率作自变量，顾客投诉次数作因变量，求出估计的回归方程，并解释回归系数的意义；

(3)检验回归系数的显著性(α=0.05)；

(4)如果航班正点率为 80%，估计顾客的投诉次数；

(5)求航班正点率为 80%时，顾客投诉次数 95%的置信区间和预测区间。

8．某汽车生产商欲了解广告费用(x)对销售量(y)的影响，收集了过去 12 年的有关数据。通过计算得到下面的有关结果：

表 1

变差来源	df	SS	MS	F	Significance F
回归					2.17E-09
残差		40158.07		—	—
总计	11	1642866.67	—	—	—

表 2

	Coefficients	标准误差	t Stat	P-value
Intercept	363.6891	62.45529	5.823191	0.000168
X Variable 1	1.420211	0.071091	19.97749	2.17E-09

要求：

(1)完成方差分析表(表 1)。

(2)汽车销售量的变差中有多少是由于广告费用的变动引起的？

(3)销售量与广告费用之间的相关系数是多少？

(4)写出估计的回归方程并解释回归系数的实际意义。

(5)检验线性关系的显著性(α=0.05)。

9．已知 12 对父子身高资料如下表：

父身高(寸)	64	63	66	65	69	62	70	66	68	67	69	71
子身高(寸)	67	66	67	66	70	66	68	65	71	67	68	70

要求：

(1)做出散点图；

(2)计算其父子身高的相关系数；

(3)估计父子身高之间的回归方程。

10．有 10 个同类企业的生产性固定资产年均价值和工业增加值资料如下：

企业编号	生产性固定资产价值(元)	工业增加值(万元)
1	320	530
2	900	1008
3	210	648
4	419	825
5	405	903
6	512	938
7	324	615
8	1230	1536
9	1052	1249
10	1125	1524
合计	6497	9776

根据资料：

(1)计算相关系数，说明两变量相关的方向和程度；

(2)估计直线回归方程，指出方程参数的经济意义；

(3)计算估计标准误；

(4)估计生产性固定资产(自变量)为 1100 万元时，工业增加值(因变量)的可能值。

11．某商场资料如下：

年份	商场销售量(千件)	价格(元/件)
2011	4	59
2012	8	54
2013	7	56
2014	9	57
2015	10	53
2016	8	57

要求：

(1)定量判断商场销售量和价格间的相关系数；

(2)用最小二乘法建立线性回归方程，并说明回归系数的经济含义；

(3)计算估计标准误差。

第四部分　时间数列分析与预测

时间数列数据是统计分析中常见的一类数据。利用时间数列数据进行统计分析不仅可以分析、了解研究对象过去变化的轨迹、特征和规律，而且可以在此基础上对研究对象的未来发展趋势进行预测分析。以下两章内容分别介绍对时间数列数据进行描述性分析和发展趋势预测分析的基本方法。

第十三章　时间数列的描述性分析

第一节　时间数列及其种类

一、时间数列

时间数列又称为时间序列，它是把不同时间上的同一指标数据按时间先后顺序排列所形成的数列。表 13-1 中包含有四个时间数列：企业月职工工资总额、年末职工人数、年轻职工所占比重和人均日生产产品数时间数列。

表 13-1　某企业职工工资总额等时间数列

	2009	2010	2011	2012	2013	2014	2015
企业月职工工资总额(万元)	9405	9926	9875	10 656	11 830	13 161	14 050
年末职工人数(人)	9500	9520	9400	9440	9510	9490	8440
年轻职工所占比重(%)	76.6	73.2	72.1	74.5	75.1	73.5	76
人均日生产产品数(件/日)	42	48	40	41	46	52	51

编制时间数列对于描述研究对象的发展变化过程，反映客观现象的发展趋势和发展速度，探索社会经济与自然现象发展变化的规律性和预测未来发展方向具有重要的意义。

二、编制时间数列的基本原则

保证指标值之间的可比性，这是编制时间数列的基本原则。可比性的具体要求主要包括以下几方面。

第一，同一时间数列的数据，其所属时间长短及数据之间的间隔长度具有可比性。一方面，由于有些指标值的大小与其所属的时间长短有直接关系，如销售额指标，时间越长，该指标数值越大，因此各指标数值所属的时间长短应当一致。另一方面，为更准确地研究现象发展变化的动态特征、规律与趋势，要求各指标数值之间的时点间隔期尽量相同。

第二，不同时期的数据，其核算范围应当一致。数据的大小与被研究现象所属总体空间范围有直接关系，当所研究范围发生了变化后，应当对前后期各数据的范围进行相应的调整，以保证核算范围的可比性。

第三，不同时期的数据所包含的内容和核算方法应具有可比性。随着时间的变化，有些统计指标，尤其是社会经济指标的内容、核算方法、计算方法都发生了变化，这时即使指标的名称相同，但前后时期指标核算内容不一致，也需要进行调整。

第四，计算价格和计量单位应具有一致性。时间数列中同类价值指标的计算价格应统一，如果用不变价格则均用不变价格，如果用现价则均用现价，切忌有些时期用不变价，而另外一些时期用现价。

三、时间数列的种类

时间数列根据指标的表现形式不同，可分为总量指标时间数列、相对指标时间数列和平均指标时间数列。例如，表 13-1 中企业月职工工资总额数列、年末职工人数数列属于总量数列，年轻职工所占比重数列是相对指标时间数列，人均日生产产品数数列则是平均数时间数列。

(一) 总量指标时间数列

总量指标是揭示总体数量绝对规模和水平的指标，其数值大小受总体范围及包含单位数的影响，将其按时间先后顺序排列而成的时间数列称为总量指标时间数列，也称为绝对数时间数列。

总量指标按其指标所属于时间的特点不同，又可进一步分为时期指标和时点指标，因而总量指标时间数列又可分为时期指标时间数列和时点指标时间数列。其中时期指标表明总体在一段时间内数量发展过程的累积结果，如某种产品的产量、职工工资总额、商品销售额等，它是通过对一定时期内事物的数量进行连续登记并累计加总得到的，表 13-1 中企业月职工工资总额时间数列就是时期数列。而时点指标是反映某一时刻或某一时点上的总量水平，其数值是通过对事物在某一时点上数量的登记，将同一时点上各部分数量加总得到的。例如年末职工人数时间数列就是时点数列。根据指标的性质，时期数列反映的是一定时期内数据的总量，各个时期数据可以相加，其数值的大小与时期的长短有关；而时点数列的各指标数值不能相加，数值大小与时间间隔长短不存在依存关系。

(二) 相对指标时间数列

相对指标是两个有联系的统计指标对比得到的派生指标，其具体数值表现为相对数。相对指标时间数列是将同类的相对指标按时间先后顺序排列起来形成的数列，它可以用来反映客观现象之间数量相互关系的发展过程。由于相对指标时间数列是派生数列，因此构成相对指标的分子、分母可以是时期指标，也可以是时点指标。由于其计算的基数不一定相同，各期的相对指标之间不具有直接可加性。如表 13-1 中的年轻职工所占比重时间数列就属于相对指标时间数列。

(三) 平均指标时间数列

这是将平均指标数值按时间先后顺序排列起来形成的数列，它可以用来反映社会经济现象总体一般水平的发展变动趋势。如表 13-1 中的人均日生产产品数时间数列就属于平均指标时间数列。与相对指标时间数列相同，平均指标时间数列的各项数据也不具有直接可加性。

第二节　时间数列的水平特征分析

进行时间数列分析，首先对研究对象在一定时期内的发展变化情况进行客观描述，以反映其发展变化的基本特征，并对其特征进行基本的测度与分析。一般经常计算的、用于测度时间数列特征的指标包括两大类：水平指标和速度指标。其中水平指标包括发展水平、平均发展水平、增长量和平均增长量；速度指标包括发展速度、增长速度、平均发展速度和平均增长速度。

一、发展水平

所谓发展水平就是指时间数列中的每一期的指标数值，反映某一现象在各个时期或时点上所达到的规模和水平，在表 13-1 中的所有各期数据均为各指标的发展水平。在对比不同

时间的发展水平时，我们所关注的、要重点研究的时间上的发展水平称为报告期水平，而作为对比基础的发展水平称为基期水平。

二、平均发展水平

时间数列的平均发展水平又称为序时平均数或动态平均数，它是将指标在不同时间上的数量差异抽象化，从动态上反映指标在一段时间内达到的一般发展水平，如 2000 年至 2016 年间的平均收入水平。平均发展水平可以用来消除某一指标在短时期内波动的影响，便于广泛地对比及观察现象的总发展态势，时间数列平均发展水平的计算需要考虑时间数列的类型与特征。

(一)根据总量数列计算平均发展水平

(1)时期数列平均发展水平的计算。时期数列中各个数据具有可加性，因而计算其平均发展水平时，可以采用简单算术平均法，即将各个时期的数据加总后除以时期的长度。其计算公式为：

$$\overline{y}=\frac{\sum_{t=1}^{n}y_t}{n}$$

利用表 13-1 中数据计算各年的企业月职工工资总额的平均数如下：

	2009	2010	2011	2012	2013	2014	2015
企业月职工工资总额(万元)	9405	9926	9875	10 656	11 830	13 161	14 050

$$\overline{y}=\frac{\sum_{t=1}^{n}y_t}{n}=\frac{9405+9926+\cdots+14050}{6}=11\,278.16\ (\text{万元})$$

如果数据不是逐日变动的，则可以根据整个时间内每次变动的资料，用每次变动持续的间隔长度为权数，对各时点水平加权，应用加权算术平均法计算，即：

$$\overline{y}=\frac{\sum_{t=1}^{k}y_t f_t}{\sum_{t=1}^{k}f_t}$$

式中，f_t 表示各数据持续的时间长度，即两时点间隔长度；y 为指标值。

(2)时点数列平均发展水平的计算。根据时点数列中数据的特征，平均发展水平的计算有以下两种计算方法。

第一，连续时点数列。连续时点数列是指可以拥有每一时点上的数据资料的数列，实际上把每个时点上的数据资料都搜集到几乎是不可能的，因此我们通常将“日”作为时点单位，这样对以日为间隔而编制的是点数列，就可以近似看成连续时点数列。对连续时点数列可以采用简单算术平均法计算其平均发展水平，计算公式与时期数列的平均发展水平相同。

$$\overline{y}=\frac{\sum_{t=1}^{n}y_t}{n}\quad 或\quad \overline{y}=\frac{\sum_{t=1}^{k}y_t f_t}{\sum_{t=1}^{k}f_t}$$

第二，有间隔的时点数列。在许多情况下，往往只能每隔一段时间才对资料进行一次登记，因此只掌握某些时点上的数据。例如，职工人数、产品库存、固定资产等，可能只在月末、季度末或年末统计数据，这样就形成了有间隔的时点数列。对于间隔时点数列的平均发展水平的计算，一般假定数据在两个相邻时点之间的变动是均匀的，这样就可以计算两个相邻时点上的数据的平均数作为这两个时点间隔内的一般水平或平均水平，从而将有间隔的时间数列转化为连续的时点数列，再用连续时点数列的平均发展水平的计算方法计算，其计算公式为：

$$\overline{y}=\frac{\left(\frac{y_1+y_2}{2}\right)f_1+\left(\frac{y_2+y_3}{2}\right)f_2+\cdots+\left(\frac{y_{n-1}+y_n}{2}\right)f_{n-1}}{f_1+f_2+\cdots+f_{n-1}}$$

式中，$f_1,f_2,\cdots,f_{n-1}$ 表示两时点间的长度。

例 13-1：利用表 13-1 中数据，计算企业月职工工资总额及职工人数时间数列的平均发展水平。

解：(1) 企业月职工工资总额时间数列的平均发展水平：

$$\overline{y}=\frac{\sum_{t=1}^{n}y_t}{n}=\frac{9405+9926+9876+10656+11830+13161+14050}{7}$$
$$=11271.86(\text{万元})$$

(2) 职工人数数列的平均发展水平：

$$\overline{y}=\frac{\left(\frac{9500+9520}{2}\right)\times1+\left(\frac{9520+9400}{2}\right)\times1+\cdots+\left(\frac{9490+8440}{2}\right)\times1}{7-1}$$
$$=9833.33(\text{人})$$

(二) 根据相对数时间数列和平均数时间数列计算平均发展水平

由于相对指标和平均指标是由两个有联系的指标对比的结果，因此相对指标和平均指标时间数列不能像总量数列那样直接计算平均发展水平，而是应先分别计算出分子、分母两个指标的平均发展水平，然后再对比得到相对指标或平均指标时间数列的平均发展水平，其计算公式是：

$$\overline{y}=\frac{\overline{a}}{\overline{b}}$$

式中，$\overline{a}$、$\overline{b}$ 分别表示分子和分母指标的平均发展水平。

例 13-2：已知某企业第一季度的商品销售额、库存额统计资料如表 13-2 所示，试计算该企业第一季度各月的商品流转次数和第一季度平均商品流转次数。

表 13-2　商品销售额与月初商品库存额数据

月份	1	2	3	4
商品销售额 a(万元)	120	143	289	—
月初商品库存额 b(万元)	50	70	60	110

解：(1) 第一季度各月商品流转次数：

$$\text{商品流转次数}=\frac{\text{商品销售额}}{\text{平均库存额}}$$

一月：$$\text{商品流转次数}=\frac{120}{(50+70)/2}=2(\text{次})$$

二月：$$\text{商品流转次数}=\frac{143}{(70+60)/2}=2.2(\text{次})$$

三月：$$\text{商品流转次数}=\frac{289}{(60+110)/2}=3.4(\text{次})$$

(2) 第一季度平均商品流转次数：

$$\overline{y}=\frac{\overline{a}}{\overline{b}}=\frac{(120+143+289)/3}{((50+70)/2+(70+60)/2+(60+110)/2)/3}=2.6(\text{次})$$

三、增长量

增长量是指时间数列中两个不同时期发展水平之差，由于计算时采用的基期不同，增长量可分为逐期增长量和累计增长量。其中逐期增长量是说明报告期比前一时期增长的绝对数量，即报告期水平与前一期水平之差。而累计增长量是报告期水平与某一固定时期水平之差，说明本期比某一固定基期增长的绝对数量，即在某一段较长时期内总的增量。其计算公式是：

$$\text{逐期增长量}=\text{报告期水平}-\text{前一期水平}=y_t-y_{t-1}$$

$$\text{累计增长量}=\text{报告期水平}-\text{某一固定时期水平}=y_t-y_0$$

从计算公式上可以看到，逐期增长量与累计增长量之间的关系是，逐期增长量之和等于相应时期的累计增长量，即：

$$y_t-y_0=\sum_{t=1}^{n}(y_t-y_{t-1})$$

利用表 13-1 数据资料计算企业月职工工资总额的增长量指标如表 13-3 所示。

表 13-3　企业月职工工资总额的增长量

	2009	2010	2011	2012	2013	2014	2015
企业月职工工资总额（万元）	9405	9926	9875	10 656	11 830	13 161	14 050
逐期增长量（万元）	—	521	–51	781	1174	1331	889
累计增长量（万元）	—	521	470	1251	2425	3756	4645

在实际工作中，有时为了消除季节变动的影响，针对各年的月度或季度数据资料计算增长量时还往往可以计算年距增长量，即：

$$\text{年距增长量}=\text{报告期某月(或某季)水平}-\text{基年同月(或同季)}$$

四、平均增长量

平均增长量是一定时期内平均每期增加(或减少)的绝对数量。一般用简单算术平均法计算，计算公式是：

$$平均增长量=\frac{逐期增长量之和}{逐期增长量的项数}=\frac{累计增长量}{数列项数-1}$$

根据表 13-3 的数据及计算，可以计算企业月职工工资总额的平均增长量如下：

$$平均增长量=\frac{逐期增长量之和}{逐期增长量的项数}=\frac{4645}{6}=774.17(万元)$$

第三节　时间数列的速度特征分析

一、发展速度

发展速度是两个不同时期发展水平的对比的结果，反映研究对象发展程度的动态相对指标，其计算公式是：

$$发展速度=\frac{报告期水平}{基期水平}$$

发展速度一般用百分数表示，有时也可以用倍数表示。由于对比的基期可以用前一期，也可以用固定时期，因而发展速度可分为环比发展速度和定基发展速度。其中，环比发展速度反映了现象逐期发展的变动程度，而定基发展速度表明了现象在较长时期内总的发展速度，也称总速度。其计算公式如下：

$$环比发展速度=\frac{报告期水平}{前一期水平}=\frac{y_t}{y_{t-1}}$$

$$定基发展速度=\frac{报告期水平}{某一固定时期水平}=\frac{y_t}{y_0}$$

利用表 13-1 数据计算出企业月职工工资总额的环比发展速度和定基发展速度如表 13-4 所示。

表 13-4　企业月职工工资总额发展速度

	2009	2010	2011	2012	2013	2014	2015
企业月职工工资总额(万元)	9405	9926	9875	10 656	11 830	13 161	14 050
环比发展速度(%)	—	105.54	99.49	107.91	111.02	111.25	106.75
定基发展速度(%)	—	105.54	105.00	113.30	125.78	139.94	149.39

从两种速度的计算公式中可以看到环比发展速度与定基发展速度之间存在如下关系。

第一，环比发展速度的连乘积等于对应时期的定基发展速度，即：

$$\frac{y_t}{y_0}=\frac{y_1}{y_0}\times\frac{y_2}{y_1}\times\cdots\times\frac{y_t}{y_{t-1}}$$

第二，相邻时期的两个定基发展速度相除，等于相应的环比发展速度，即：

$$\frac{y_t}{y_0}\div\frac{y_{t-1}}{y_0}=\frac{y_t}{y_{t-1}}$$

此外，对于具有季节变动的一些社会经济现象，为了消除季节变动的影响，可以计算年

距发展速度，用以说明本期发展水平与去年同期发展水平对比的发展程度。其计算公式是：

$$\text{年距发展速度}=\frac{\text{本年某月(季)发展水平}}{\text{去年同月(季)发展水平}}$$

二、增长速度

增长速度是增长量与基期水平对比的结果，反映现象增长程度的动态相对指标，其计算公式是：

$$\text{增长速度}=\frac{\text{增长量}}{\text{基期发展水平}}=\text{发展速度}-1$$

增长速度与发展速度相同，由于计算增长量(或发展速度)采用的基期不同，因而可分为环比增长速度和定基增长速度。环比增长速度是逐期增长量与前一时期发展水平对比的结果，反映现象的逐期增长程度；定基增长速度是累计增长量与某一固定时期发展水平对比的相对数，反映现象在较长时期内总的增长程度，其公式为：

$$\text{环比增长速度}=\frac{\text{逐期增长量}}{\text{前一期发展水平}}=\text{环比发展速度}-1$$

$$\text{定基增长速度}=\frac{\text{累计增长量}}{\text{某一固定时期发展水平}}=\text{定基发展速度}-1$$

利用表 13-1 中数据资料计算企业每月职工工资总额增长速度如表 13-5 所示。

表 13-5　企业每月职工工资总额增长速度

	2009	2010	2011	2012	2013	2014	2015
企业月职工工资总额(万元)	9405	9926	9875	10 656	11 830	13 161	14 050
环比增长速度(%)	—	5.54	–0.51	7.91	11.02	11.25	6.75
定基增长速度(%)	—	5.54	5.00	13.30	25.78	39.94	49.39

与发展速度不同，环比增长速度与定基增长速度之间不存在着直接的数量关系，即：

$$\text{定基增长速度}\neq\text{环比增长速度之乘积}$$

与年距发展速度相对应，年距增长速度是年距增长量与前一年同期水平对比的结果，也可以用年距发展速度–1，即：

$$\text{年距增长速度}=\frac{\text{年距增长量}}{\text{去年同期发展水平}}=\text{年距发展速度}-1$$

三、平均发展速度与平均增长速度

平均发展速度是一定时期内各个环比发展速度的平均数，反映现象在一定时期中逐期平均发展变化的程度。平均增长速度则是一定时期内各环比增长速度的一般水平，反映现象在一个较长时期中逐期平均增长变化的程度，但它不是根据各环比增长速度直接计算的，而是根据平均发展速度计算的，即：

$$\text{平均增长速度}=\text{平均发展速度}-1$$

平均增长速度在实际工作中具有重要的作用，它通常用来对比不同阶段、不同时期、不同国家或地区同类现象发展变化的情况，但其计算往往是通过平均发展速度进行的。

平均发展速度的计算方法有水平法(又称几何平均法)和累计法(又称方程式法)。

(一)水平法

水平法计算平均发展速度，是计算各环比发展速度的几何平均数，即：

$$\bar{x}=\sqrt[n]{x_1x_2\cdots x_n}=\sqrt[n]{\frac{y_1}{y_0}\times\frac{y_2}{y_1}\times\ldots\times\frac{y_n}{y_{n-1}}}=\sqrt[n]{\frac{y_n}{y_0}}$$

利用表 13-1 数据资料计算企业月职工工资总额平均发展速度得到：

$$\bar{x}=\sqrt[6]{\frac{14050}{9405}}=106.92\%$$

平均增长速度为：　　平均发展速度 $-100=6.92\%$

(二)累计法

这种方法要求，按平均发展速度计算出来的各期水平累计总和与各年实际所具有的发展水平累计总和相等，即：

$$y_0\bar{x}+y_0\bar{x}^2+\cdots+y_0\bar{x}^{n-1}+y_0\bar{x}^n=y_1+\cdots+y_{n-1}+y_n$$

这样：

$$y_0(\bar{x}+\bar{x}^2+\cdots+\bar{x}^{n-1}+\bar{x}^n)=\sum_{t=1}^{n}y_t$$

$$\bar{x}+\bar{x}^2+\cdots+\bar{x}^{n-1}+\bar{x}^n=\frac{\sum_{t=1}^{n}y_t}{y_0}$$

解此方程即可得到平均发展速度 $\bar{x}$。

计算平均发展速度的两种方法，各有不同的出发点和应用条件，当然计算结果也不相同。其中，水平法侧重于考察期末发展水平，它不反映中间各项水平的变化，所以在计算平均发展速度时，必须对间隔期内的各期经济情况进行分析。如果中间各期发展水平忽高忽低或最末水平受特殊因素的影响而过高或过低时，运用水平法计算出的平均发展速度就没有代表性。而累计法则侧重于考察整个时期中各年发展水平的总和，因此利用这种方法计算出的发展水平，决定于间隔期内中间各个时期的变化情况。

例 13-3：以下是某企业 2007—2015 年生产产品产量资料(如表 13-6 所示)，要求根据数据资料计算逐期增长量、累计增长量、环比发展速度、定基发展速度、水平法平均发展速度、环比增长速度、定基增长速度和平均增长速度。

表 13-6　2007—2015 年产品产量资料

	2007	2008	2009	2010	2011	2012	2013	2014	2015
产量(吨)	1089	1156	1242	1285	1516	1823	1915	1980	2015

解：(1)根据上述资料计算增长量、发展速度和增长速度，如下表所示。

		2007	2008	2009	2010	2011	2012	2013	2014	2015
增长量(吨)	逐期	—	67	86	43	231	307	92	65	35
	累计	—	67	153	196	427	734	826	891	926
发展速度(%)	环比	—	106.15	107.44	103.46	117.98	120.25	105.05	103.39	101.77
	定基	—	106.15	114.05	118.00	139.21	167.40	175.85	181.82	185.03
增长速度(%)	环比	—	6.15	7.44	3.46	17.98	20.25	5.05	3.39	1.77
	定基	—	6.15	14.05	18.00	39.21	67.40	75.85	81.82	85.03

(2)平均发展速度。

用水平法计算平均发展速度为：

$$\bar{x}=\sqrt[n]{\frac{y_n}{y_0}}=\sqrt[8]{\frac{2015}{1089}}=108\%$$

(3)平均增长速度=108%–100%=8%。

思考与练习

1．简述时间数列及其构成要素。

2．编制时间数列应注意的基本问题有哪些？

3．简述时间数列的基本种类和特点。

4．简述时期数列和时点数列的区别。

5．举例说明时期数列和时点数列的特点。

6．对时间数列进行描述的基本指标都有哪些？

7．已知企业第一季度的商品销售额、人数的统计资料，试计算该企业第一季度各月的人均销售额和第一季度人均月商品销售额。

月份	1	2	3	4
商品销售额 a(万元)	120	143	289	—
月初人数 b(人)	50	70	60	110

8．以下是某企业 2007—2015 年生产产品产量资料，要求根据数据资料计算逐期增长量、累计增长量、环比发展速度、定基发展速度、水平法平均发展速度、环比增长速度、定基增长速度和平均增长速度。

	2007	2008	2009	2010	2011	2012	2013	2014	2015
产量(吨)	1005	1150	1240	1280	1520	1830	1920	1990	2020

第十四章　时间数列的构成与预测

第一节　时间数列的构成要素与模式

影响社会经济现象发展变化的因素很多，有些因素是属于基本因素，它对事物的发展变化起决定性作用，会使事物的发展呈现出一定的规律性；有些因素是属于偶然的非基本因素，对事物的发展变化不会起到决定的作用，表现出一种不规则的波动。为了对事物的未来进行预测，就需要了解事物在过去的一段时间里是如何变化的，掌握事物发展变化的形态、趋势与规律，进而建立适当的预测模型。

一、时间数列的构成要素

导致时间数列形成的原因很多，但从另一角度，我们可以将时间数列分解为四个基本要素，即：长期趋势、季节变动、循环变动、不规则波动。

(一) 长期趋势

长期趋势是指客观现象在一个相当长的时期内，受某种稳定性因素影响所呈现出的上升或下降的趋势。例如，由于投入要素数量的增多、质量的提高、技术的进步、需求上升等因素，我国的国内生产总值表现出逐年稳步上升的趋势即是长期趋势。长期趋势可能是线性的(如图 14-1 左所示)，也可能是非线性的(如图 14-1 右所示)。

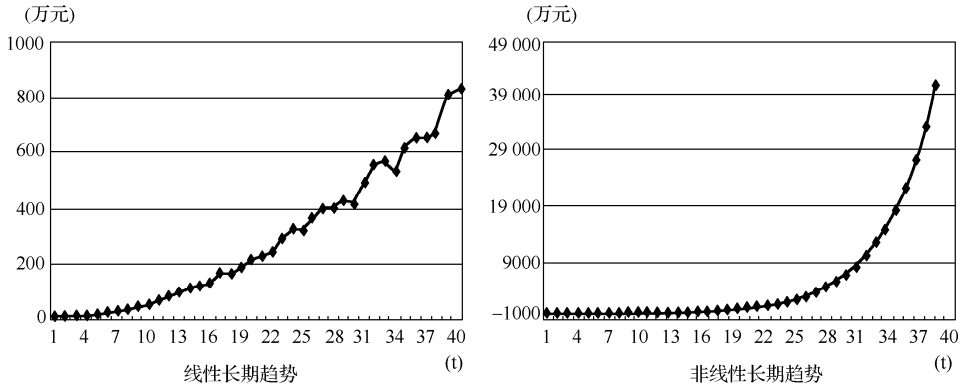


图 14-1　长期趋势

(二) 季节变动

季节变动是时间数列受季节因素的影响，在一定时期内随季节变化呈现出来的一种周期性的波动。如工业企业产品订单会受到季节因素的影响形成季节性的波动；零售企业商品零售额也会随着季节变化表现出明显的季节波动，像十一、元旦、春节期间的销售额明显高于其他时间的销售额；夏季是空调的销售旺季，明显高于其他季节性的销售；铁路和航空客运

在节假日会迎来客流的高峰等。这样的订单数、零售额、客流量等时间数列都包含明显的季节变动，其图形类似于图 14-2 的形状。

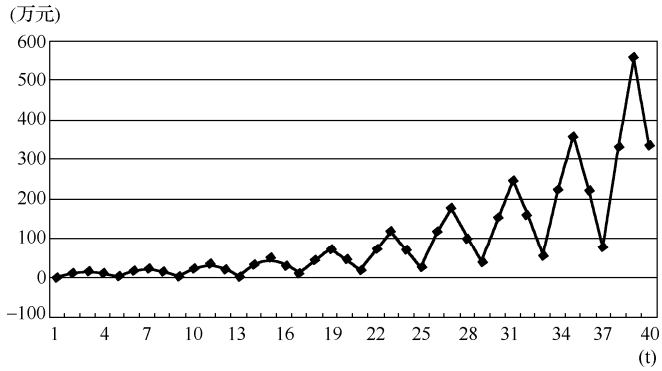


图 14-2　包含有季节变动的时间数列

(三)循环变动

循环变动也是一种周期性的波动，但其是非固定周期长度且周期相对较长的一种周期性波动，即事物以若干年为周期的涨落起伏的变动。在经济研究中经济周期、景气周期即是一种循环变动。循环波动的周期可能会持续一段时间，但与长期趋势不同，它不是朝着一个方向持续变动，而是涨落起伏；它与季节变动也不同，循环变动周期性相对较长，且不固定，而季节变动周期较短，一般短于一年。

(四)不规则波动

不规则波动是指客观现象由于突发事件或偶然因素引起的无规律性的变动，也称为随机波动。

一个时间数列可能由一种要素构成，也可能同时包含有多种构成要素，图 14-3 就包含有季节变动、长期趋势和不规则波动三个要素。

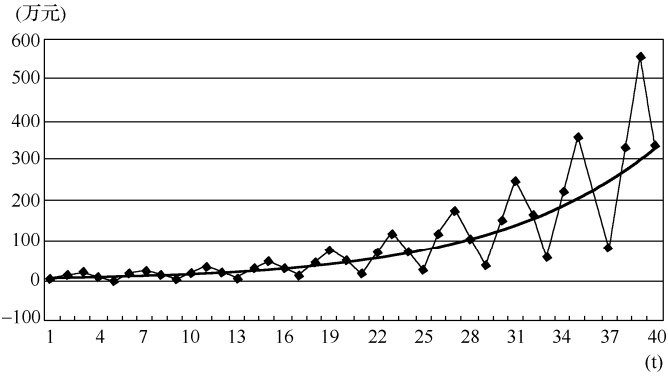


图 14-3　包含有季节波动、长期趋势和不规则波动的时间数列

二、时间数列的构成模式

以上四种因素的变化构成了事物在一定时期内的变动。在分析时间数列时，首先要明确

这四种类型因素变动的构成形式，即它们是如何结合及相互作用的。把这些构成要素和时间数列的关系用一定的数学关系表示，就构成了时间数列影响因素分解模型。一般常用的数学模型有加法模型和乘法模型。

(1)加法模型假定四种变动因素是相互独立的，则时间数列各期发展水平是各个影响因素相加的总和。其结构是：

$$Y = T + S + C + I$$

式中，Y 表示长期趋势；S 表示季节变动；C 表示循环变动；I 表示不规则变动。

(2)乘法模型则假定四种构成要素存在着某种相互影响的关系，互不独立，因此时间数列各期发展水平是各个构成要素相乘之积。其基本结构是：

$$Y = T \times S \times C \times I$$

当然，有时一个复杂的时间数列在表面上看不出其基本构成要素，但是将其进行多种分解后其变动趋势就比较明显了。

例如，某商场某种品牌化妆品的销售额，如表 14-1 所示。

表 14-1　某种品牌化妆品的销售额

年	月	销售额合计(万元)	其中			
			保湿 x1(万元)	美白 x2(万元)	抗皱 x3(万元)	防晒 x4(万元)
2013	1	25.424	9.231	4.563	10.08	1.55
	2	16.539	7.452	4.322	3.07	1.695
	3	29.653	6.86	14.561	3.11	5.122
	4	30.87	5.21	16.23	2.98	6.45
	5	28.983	3.851	14.622	2.95	7.56
	6	24.415	3.424	12.411	3.03	5.55
	7	25.317	2.845	13.88	3.05	5.542
	8	25.22	1.33	16.42	2.85	4.62
	9	25.304	12.56	8.339	3.21	1.195
	10	27.152	16.85	6.212	3.07	1.02
	11	26.612	17.03	5.022	3.04	1.52
	12	26.382	11.76	3.332	9.6	1.69
2014	1	28.164	12.413	3.447	10.652	1.652
	2	23.588	11.335	3.672	6.749	1.832
	3	32.93	11.225	14.621	2.563	4.521
	4	34.745	7.512	17.443	3.115	6.675
	5	31.32	5.567	15.562	2.775	7.416
	6	22.664	3.325	11.501	2.618	5.22
	7	22.866	2.526	13.972	1.198	5.17
	8	23.682	4.338	12.712	2.052	4.58
	9	27.493	13.892	9.765	2.164	1.672
	10	25.215	15.432	6.523	2.149	1.111
	11	25.993	16.113	5.114	3.454	1.312
	12	29.174	12.974	3.756	10.839	1.605
2015	1	29.63	14.031	4.213	9.857	1.529
	2	24.584	10.965	6.021	5.864	1.734
	3	33.513	10.845	13.528	3.511	5.629
	4	34.26	6.394	16.924	4.021	6.921
	5	30.184	4.859	13.526	3.265	8.534
	6	24.296	4.335	12.117	2.548	5.296
	7	25.337	3.528	13.528	2.781	5.5
	8	26.633	6.531	11.968	3.023	5.111
	9	24.884	12.358	7.365	3.627	1.534

续表

年	月	销售额合计(万元)	其中			
			保湿 x1(万元)	美白 x2(万元)	抗皱 x3(万元)	防晒 x4(万元)
2015	10	26.123	15.449	5.857	3.512	1.305
	11	27.11	16.387	4.936	4.259	1.528
	12	31.133	15.326	3.586	10.597	1.624

将该品牌化妆品销售额合计的时间数列数据绘制成折线图，如图 14-4 所示。

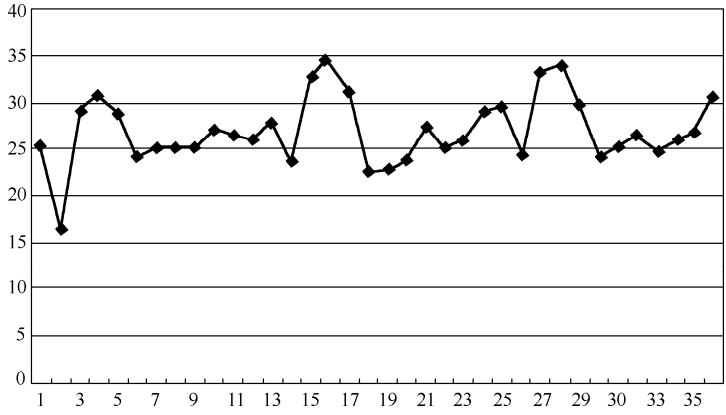


图 14-4　某品牌化妆品的销售额(单位：万元)

从图 14-4 中不易看出其明显的变化规律特征，但若将不同类型产品分别进行分析，绘制时间数列折线图，就可以看到比较明显的销售规律(如图 14-5 所示)，该品牌产品销售总额基本保持稳定，但各种产品的销售存在着明显的季节变动。

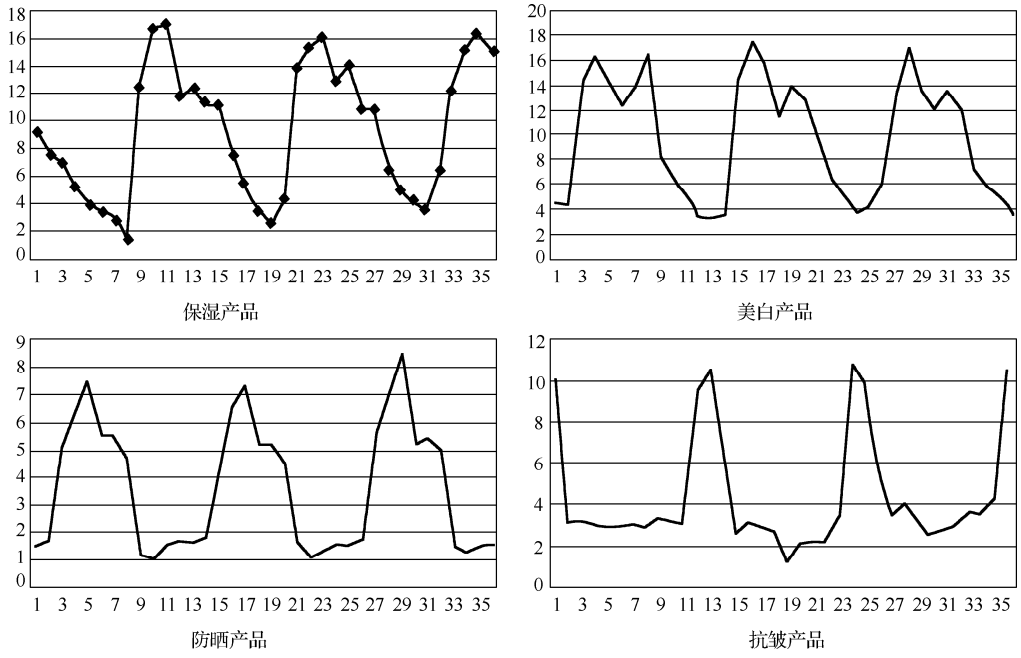


图 14-5　分类型化妆品的销售额(单位：万元)

第二节　时间数列的长期趋势与预测分析

长期趋势是指客观现象受某种普遍的、持续的、决定性的、稳定性因素的影响，各期发展水平在相当长的时间内沿着一定的方向上升或下降的势态。因此，长期趋势是时间数列变动中最基本的形式。研究长期趋势，有助于认识客观现象的变动规律，可以为预测事物未来的发展情况提供依据。

一、长期趋势的确定——时间数列的修匀

时间数列长期趋势的确定，主要是对时间数列进行修匀，即剔除或削弱波动，突出长期趋势的特征。修匀的方法主要有时距扩大法、移动平均法、趋势线配合法等，其主要目的是通过将时间数列中的其他构成要素剔除，直接显示其最主要的长期趋势。

（一）时距扩大法

时距扩大法是把原来时间数列中各数据的时间间隔扩大，求各数据的和或平均数，得出较长时间的时距资料，组成新的时间数列，用以消除由于时距较短受偶然因素影响所引起的波动。这样，经过整理后的时间数列就可以清楚地反映出数据变动的总趋势。

例 14-1：某商品连续四年的季度销售量资料如表 14-2 所示。

表 14-2　季度销售量资料

	2012				2013				2014				2015			
季度	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
销售量（万件）	13	18	5	8	14	18	6	10	16	22	8	12	19	25	15	17

从表 14-2 中数据可以看到，2012—2015 年各季度的销售量由于受多种因素的影响，增长趋势不够明显。但是如果将其时间扩大为年，则可整理出新的时间数列数据，如表 14-3 所示。

表 14-3　年度销售量资料

年度	2012	2013	2014	2015
销售量（万件）	44	48	58	76
季平均销售量	11	12	14．5	19

表 14-3 中的销售量数据是时距扩大后按年计算的总数，季平均销售量是时距扩大后计算的平均数，从处理后的数据可以明显地看到该商品的销售量有明显的上升趋势，即长期趋势。

运用时距扩大法修匀时间数列时，应该使各时期扩大的时距长短保持一致，否则难以比较。在确定时距时，时距的大小要适中。如果时距过大，整理出的新的时间数列数据太少，现象的发展的具体变化过程被掩盖；如果时距过小，则偶然因素或季节波动不易消除，就不能达到反映现象长期的发展趋势的目的。一般来说，扩大时距应与现象的变化周期相一致，否则也会影响对发展趋势的判断与分析。

（二）移动平均法

移动平均法是采取逐项依次递移的方法将时间数列的时距扩大，计算扩大时距后的序时平均数，形成一个新的时间数列。在这一新的数列中，由于短期起作用的偶然因素的影响已经削弱，甚至已被剔除，从而显示出现象发展的基本趋势。

例 4-2：某企业各期产品产量数据资料如表 14-4 第 2 列所示，根据产值数据资料计算移动平均数。

表 14-4　移动平均计算表

	总产值(万元)	三项移动平均	四项移动平均	移正平均
(1)	(2)	(3)	(4)	(5)
1	81	—	—	—
2	70	78.33	—	—
3	84	78.33	79.00	80.38
4	81	85.67	81.75	85.13
5	92	90.00	88.50	89.63
6	97	94.00	90.75	92.88
7	93	96.00	95.00	95.75
8	98	96.33	96.50	97.25
9	98	99.67	98.00	98.88
10	103	100.33	99.75	101.13
11	100	104.00	102.50	—
12	109	—	—	—

移动平均数的计算需要确定计算平均数的项数，移动平均数的项数有奇数项和偶数项。其中奇数项移动平均求的平均数，应对准所平均时期的中间时期，如利用第 1、2、3 期数据计算的移动平均数 78.33 应置于第 2 期，而用第 3、4、5 期数据计算的移动平均数 85.67 应置于第 4 期，等等。而偶数项移动平均计算得到的移动平均数，应置于所平均时期的中间两项之间，如计算四项移动平均数时，如用第 1、2、3、4 期数据计算的平均数 79 应置于第 2、3 期中间，以此类推。这样组成的新数列中，每个数值都错后半期，这时可采用移正平均的方式，即利用四项移动平均数再计算一次二项移动平均，使之与具体的时间相对应，计算结果见表 14-4 中的(3)、(4)、(5)列。

从图 14-6 和图 14-7 中可以看到，移动平均后的时间数列对原时间数列做了修匀，去掉了部分波动，因而比原时间数列要平滑一些。采用移动平均法所得的新数列项数比原数列项数要少。一般来讲，被平均的项数越多，修匀的作用越大，数据就越平滑，而所得到的移动平均数就越少。一般情况下，若要消除或降低随机波动(不规则波动)，可以计算奇数项移动平均数，移动平均的项数要适中，否则不利于现象的发展趋势的分析；当然，如果数列存在的周期性变动是季节变动，则为去掉该波动可以用周期的长度作为被平均的项数。

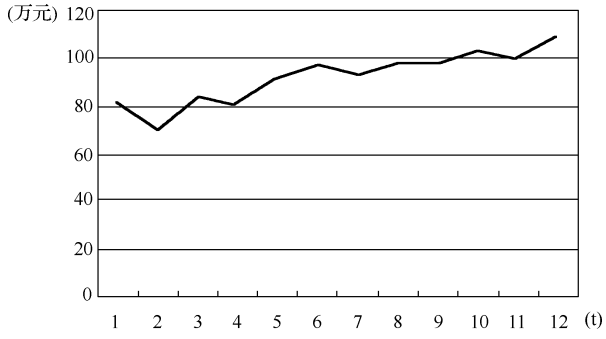


图 14-6　原时间数列图

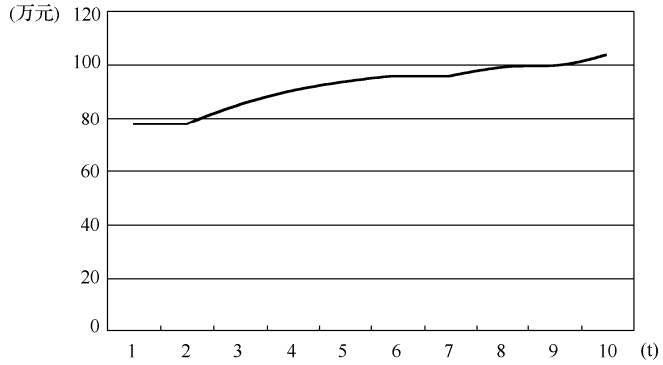


图 14-7　三项移动平均数列图

二、长期趋势模型的建立——趋势线配合

趋势线配合法是依据数学模型给时间数列配合一条较为理想的趋势线，然后据此计算其趋势值。在建立趋势线方程之前，首先要确定趋势线的形态，最常用的确定方法是画折线图。若散点图属直线趋势形态，则可配合直线方程；若为曲线形态，则可配合曲线方程。此外，还可以根据动态分析指标判断。以下介绍线性及非线性趋势两类模型的建立。

(一)线性趋势模型

线性趋势模型是：　　$T = a + bt$

对于线性趋势模型，只要估计出模型中的参数 a、b 的值，并将时间 t 代入模型中，即可计算出各期的趋势值及未来某一时间的预测值。线性趋势模型参数的最常用的估计方法是最小二乘法。

最小二乘法，也称为最小平方法。这种方法的数学依据是实际值 Y_t 与趋势值 $\hat{T}_t$ 的离差平方和为最小，即 $\sum (Y_t - \hat{T}_t)^2 = \sum (Y_t - a - bt)^2$ 为最小值。具体方法是，采用偏导数求极值的方法建立二元联立方程，求解 a、b 的值，其计算公式是：

$$\begin{cases} b = \dfrac{n\sum tY - \sum t \sum Y}{n\sum t^2 - (\sum t)^2} \\ a = \overline{Y} - b\overline{t} \end{cases}$$

例 14-3：已知某时间数列数据如表 14-5 所示。

表 14-5　某时间数列数据

t	1	2	3	4	5	6	7	8	9	10	11	12
总产值(万元)	81	70	84	81	92	97	93	98	98	103	100	109

绘制时间数列散点图可以看到该时间数列总体呈现出线性趋势(略)，故可配合直线趋势模型 $\hat{Y} = a + bt$。根据已知数据资料计算参数的过程如下(如表 14-6 所示)。

表 14-6　计算表

年份	t	总产值 Y	tY	t^2
1998	1	81	81	1
1999	2	70	140	4

续表

年份	t	总产值 Y	tY	t^2
2000	3	84	252	9
2001	4	81	324	16
2002	5	92	460	25
2003	6	97	582	36
2004	7	93	651	49
2005	8	98	784	64
2006	9	98	882	81
2007	10	103	1030	100
2008	11	100	1100	121
2009	12	109	1308	144
合计	78	1106	7594	650

将数据代入参数的估计公式得到：

$$\begin{cases} b = 2.83 \\ a = 73.76 \end{cases}$$

则：

$$\hat{T} = 73.76 + 2.83t$$

若要预测第 13 期数值，则将 t=13 代入模型中，即可得到：

$$\hat{T}_{13} = 73.76 + 2.83 \times 13 = 110.55 \text{（万元）}$$

（二）非线性趋势模型

实际工作中，我们会发现很多现象的发展变化趋势并不总是表现为直线趋势，更多的情况可能会呈现出某种曲线变动趋势。当然，曲线有许多不同的种类与形态，常用的非线性曲线趋势模型主要有以下几个。

1. 二次曲线趋势模型

在时间数列中，当各期发展水平的二次增长量大致相同，且散点图围绕二次抛物线趋势波动时，可以配合二次曲线趋势模型进行长期趋势分析及预测。

二次曲线趋势模型是：

$$T = a + bt + ct^2$$

其中，a、b、c 为模型的参数。

二次曲线趋势模型最常用的参数估计的方法是最小二乘法，其基本原理是：求实际值 Y_t 与趋势值 $\hat{T}_t$ 的离差平方和为最小，即 $\sum (Y_t - \hat{T}_t)^2 = \sum (Y_t - a - bt - ct^2)^2$ 为最小值时的 a、b、c。

对此式求偏导数得到三个标准方程式如下：

$$\begin{cases} \sum Y = na + b\sum t + c\sum t^2 \\ \sum tY = a\sum t + b\sum t^2 + c\sum t^3 \\ \sum t^2 Y = a\sum t^2 + b\sum t^3 + c\sum t^4 \end{cases}$$

据此三元联立方程组，可求解出 a、b、c 三个参数，即完成二次曲线趋势模型的配合。

2. 指数曲线趋势模型

指数曲线模型用于描述以几何级数递增或递减的现象，即时间数列的观察值 Y_t 按指数规律变化，或者说在较长时期内时间数列的观察值的环比发展速度或环比增长速度比较稳定。指数曲线趋势模型是：

$$T_t = ab^t$$

其中，a、b 为模型参数。

至于指数曲线模型参数的估计方法，可以通过数学变换的方法先将模型线性化，然后再采用最小二乘法计算参数的估计值。

在指数曲线模型等号两边同时取对数，得到：$\lg T_t = \lg a + t\lg b$。

然后进行数据变换，令 $T' = \lg T_t$，$A = \lg a$，$B = \lg b$。

则：

$$T' = A + Bt$$

对此模型可以采用最小二乘法估计参数 A 和 B，再求其反对数得到 a、b。

此外，更简单但结果比较粗糙的方法是将 a 作为基期水平，b 看作平均发展速度，t 为时间，只要用水平法计算出时间数列的平均发展速度后，就可得到指数曲线模型。

例 14-4：已知某企业近 10 年产值数据资料如表 14-7 所示，求配合趋势线模型。

表 14-7　数据及计算表

年份	t	产值 Y（万元）	$\lg T$	$t\lg T$	t^2
2000	1	1395	3.144574	3.144574	1
2001	2	1515	3.180413	6.360825	4
2002	3	1729	3.237795	9.713385	9
2003	4	1916	3.282396	13.12958	16
2004	5	2146	3.33163	16.65815	25
2005	6	2393	3.378943	20.27366	36
2006	7	2682	3.428459	23.99921	49
2007	8	2999	3.476976	27.81581	64
2008	9	3374	3.528145	31.75331	81
2009	10	3812	3.581153	35.81153	100
合计	55	—	33.57048	188.66	385

首先，根据表 14-7 给出的产值数据资料，得到曲线图，如图 14-8 所示，可考虑配合指数曲线。

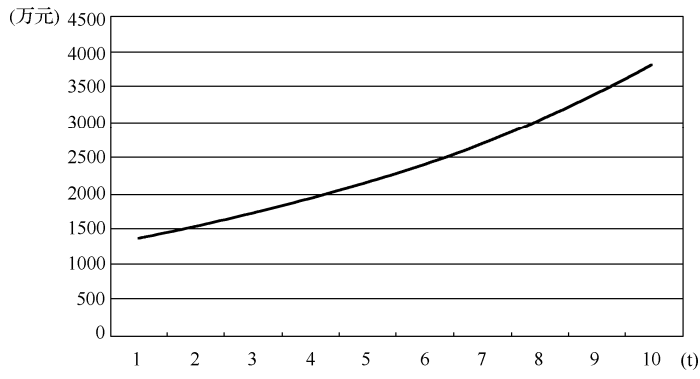


图 14-8　曲线图

然后，进行模型的变换，令 $T'=\lg Y_t$，$A=\lg a$，$B=\lg b$；

采用最小二乘法估计 A 和 B，计算过程见表 14-7。

最后得到：

$$\begin{cases} B=\dfrac{n\sum t\lg Y-\sum t\sum \lg Y}{n\sum t^2-(\sum t)^2}=0.04876 \\ A=\lg \overline{Y}-B\overline{t}=3.089 \end{cases}$$

求 A 和 B 的反对数得到：$a=1227.13$，$b=1.1188$。

则配合的指数曲线模型为：$\hat{T}=1227.13\times(1.1188)^t$。

3．修正指数曲线趋势模型

在指数曲线中，我们可以看到，若 $b>1$，增长率随着时间 t 的增大而增大，估计的趋势值将会趋向无穷大；若 $a>0$，$b<1$，则随着时间的延伸趋势值将会趋向于 0。而这与实际往往是不相符合的。而修正指数曲线模型则是在指数曲线的基础上增加一个常数项 K，且要求 $K>0$，$a\neq 0$，$0<b<1$。修正指数曲线模型主要用于描述这样一类现象，即初期增长迅速，随后增长率逐渐降低，最终则收敛于正的常数极限 K。生活中有许多事物的发展过程符合修正指数曲线模型的特点。例如，某种刚上市的新产品，初期销售增长可能很快，当社会拥有量达到一定程度时，其销售增长减缓，最终销售量逐渐趋向于某一稳定的水平。

修正指数曲线趋势模型为：$T=K+ab^t$，其中 K，a 和 b 为待估计参数。

修正指数曲线模型参数估计的方法比较特殊，当增长上限 K 已知时，可将时间数列的观测值减去 K 值，然后采用指数曲线参数估计的方法进行估计。时当增长上限 K 未知时，一般采用三和法进行估计。

所谓三和法，其基本思想是将时间数列观察值等分为三份，每份均有 m 个数据，将所有观测值 Y_t、t 代入修正指数曲线模型中，然后将每份数据所对应的方程相加，求解三个方程即可得到三个参数的估计值。

例 14-5：已知某产品的销售资料 Y 如表 14-8 所示，试配合修正指数曲线模型。

表 14-8　数据及计算表

t	Y(万元)	$T=K+ab^t$	局部求和得到的方程
0	150	$150=K+ab^0$	$330=2k+ ab^0+ ab^1$　(1)
1	180	$180=K+ab^1$	
2	204	$204=K+ab^2$	$428=2k+ ab^2+ab^3$　(2)
3	224	$224=K+ab^3$	
4	240	$240=K+ab^4$	$493=2k+ ab^4+ab^5$　(3)
5	253	$253=K+ab^5$	

解：将时间数列等分为三份，将各组数据代入修正指数曲线模型中，并求和(上述计算见表 14-8 计算结果)得到三个方程式。

用方程(2)－(1)得到：$98=ab^3+ab^2-ab^1-ab^0=a(b^3+b^2-b-1)$　(4)

用方程(3)－(2)得到：$65=ab^5+ab^4-ab^3-ab^2=ab^2(b^3-b^2-b-1)$　(5)

用式(5)除以式(4)得到：

$$b^2 = 0.6632 \qquad b = 0.8144$$
$$a = -160.4 \qquad k = 310.5$$

这样，配合的修正指数曲线模型为：

$$\hat{T} = 310.5 - 160.4 \times (0.8144)^t$$

利用三和法计算模型参数，也可以直接使用下面公式：

$$\begin{cases} b = \left(\dfrac{S_3 - S_2}{S_2 - S_1}\right)^{\frac{1}{m}} \\ a = (S_2 - S_1)\dfrac{b-1}{b(b^m-1)^2} \\ K = \dfrac{1}{m}\left(S_1 - \dfrac{ab(b^m-1)}{b-1}\right) \end{cases}$$

其中，S_1、S_2、S_3 分别为三组数据之和，m 为每组数据的个数。

4. 龚帕兹(Gompertz)曲线趋势模型

修正指数曲线模型考虑了现象的增长上限的特点，但未考虑曲线斜率的变化速度。而实际中有些客观现象的时间数列常常具有如下特征：初期阶段以较慢的速度逐渐增长，而后增长速度加快，在达到一定水平后，增长速度减缓，到后期逐渐收敛于某一稳定的水平。用图表示大致为一条 S 曲线(如图 14-9 所示)。在这种曲线上存在一个拐点，即增长速度由上升突变为下降的点，另外还具有一个增长的上限。配合这种变化趋势的曲线模型之一是龚帕兹曲线趋势模型。

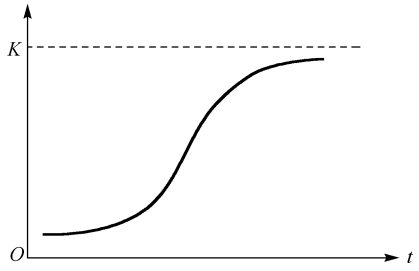


图 14-9　曲线图

龚帕兹曲线趋势模型的形式为：

$$T = K \cdot a^{b^t} \qquad (0 < a < 1, 0 < b < 1, K > 0)$$

龚帕兹曲线趋势模型的参数估计可以借助于修正指数曲线模型参数估计的方法——三和法进行，但在使用三和法前首先需要对龚帕兹曲线模型变形为修正指数曲线模型的形式。即首先对龚帕兹曲线趋势模型两边取对数得到：

$$\lg Y = \lg K + (\lg a)b^t$$

如令：

$$T' = \lg Y, \quad K' = \lg K, \quad A = \lg a$$

上式变为：

$$T' = K' + Ab^t$$

这恰好是修正指数曲线模型的形式。仿照修正指数曲线模型参数估计的方法，可得到 b、$\lg a$、$\lg K$ 的值。即：

$$\begin{cases} b=\left(\dfrac{S_3-S_2}{S_2-S_1}\right)^{\frac{1}{m}} \\ \lg a=(S_2-S_1)\dfrac{b-1}{b(b^m-1)^2} \\ \lg K=\dfrac{1}{m}\left(S_1-\dfrac{\lg a\cdot b(b^m-1)}{b-1}\right) \end{cases}$$

其中，S_1、S_2、S_3 分别为三组数据 Y 的对数 $\lg Y$ 之和，m 为每组数据个数。

最后通过求反对数得到 a 和 K。

例 14-6：某品牌 MP3 在一个地区的销售量的统计数据如表 14-9 所示，试建立龚帕兹曲线趋势模型。

表 14-9　数据及计算表

年份	t	销售量 Y(台)	lgY	局部和
2001	1	13	1.114	
2002	2	18	1.255	3.832
2003	3	29	1.462	
2004	4	48	1.681	
2005	5	52	1.716	5.242
2006	6	70	1.845	
2007	7	74	1.869	
2008	8	78	1.892	5.653
2009	9	78	1.892	

解：利用销售量数据绘制时间数列曲线图(如图 14-10 所示)，可以看到具有龚帕兹曲线趋势的基本特征，故可配合龚帕兹曲线趋势模型。

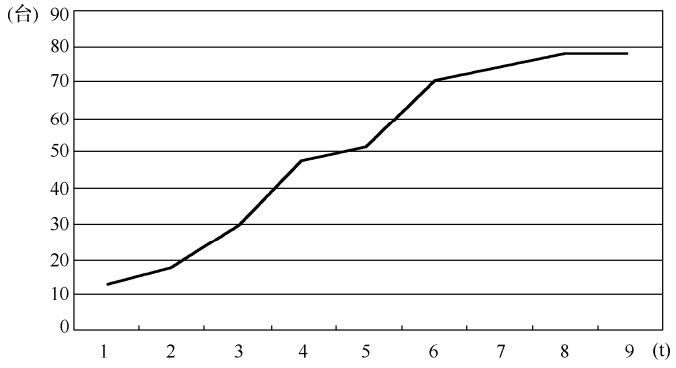


图 14-10　数列曲线图

有关计算过程，包括对时间数列观测值取对数，并进行局部求和等(如表 14-9 所示)。

根据计算公式得：

$$\begin{cases} b=\left(\dfrac{S_3-S_2}{S_2-S_1}\right)^{\frac{1}{m}}=\left(\dfrac{5.653-5.242}{5.242-3.832}\right)^{\frac{1}{3}}=0.663 \\ \lg a=(S_2-S_1)\dfrac{b-1}{b(b^m-1)^2}=(5.242-3.832)\dfrac{0.663-1}{0.663(0.663^3-1)^2}=-1.428 \\ \lg K=\dfrac{1}{m}\left(S_1-\dfrac{\lg a\cdot b(b^m-1)}{b-1}\right)=\dfrac{1}{3}\left(3.832-\dfrac{(-1.428)\times 0.663(0.663^3-1)}{0.663-1}\right)=1.941 \end{cases}$$

根据上述计算结果得到：

$$a=0.0373,\quad K=87.26$$

配合的龚帕兹曲线趋势模型为：$\hat{T}=87.26\times 0.0373^{0.663^t}$

三、长期趋势模型的选择

上面介绍了对时间数列的长期趋势配合趋势线的一般方法。但在实际应用中，对于一个具体的时间数列，应如何选择所要配合的趋势线类型是我们必须面对的一个问题。趋势线的选择是一个十分重要的问题，它直接关系到我们对现象的描述及其规律性的认识的结论。趋势线选择得不合适，不仅不能正确地描述出现象的发展变化的规律，而且还可能得出错误的结论。一般确定长期趋势模型形式的方法有以下两类。

一类是根据时间数列的观测数据绘制时间数列曲线图，观察图形的基本特征然后选择相对比较合适的、符合现象发展变化特征的趋势线。若绘制出的时间数列曲线图近似地表现为一条直线，则可配合直线趋势模型；若近似地表现为抛物线形式，则可配合二次曲线模型；若表现为 S 形曲线则可配合龚帕兹曲线趋势模型等。

第二类方法是可以根据数据本身的特点确定曲线趋势模型的类型。若观察数据的逐期增长量大致相同，则可配合直线趋势模型；若时间数列的观察数据的二次差即逐期增长量的增长量大致相同，则可配合二次曲线模型；若观察数据的环比发展速度或环比增长速度大致相同，则可配合指数曲线模型；若各观察值的一次差的环比值大体相同，则可配合修正指数曲线模型；若观察值对数的一次差的环比值大致相同，则可配合龚帕兹曲线趋势模型。

当然，最后如果对同一时间数列有多种趋势模型可考虑，则可以选择估计标准误差最小的模型作为趋势模型。估计标准误差的计算公式为：

$$S_y=\sqrt{\frac{\sum(Y-T)^2}{n-m}}$$

其中，m 为趋势方程中需要估计的参数的个数。

第三节　时间数列的季节变动分析

在现实生活中，季节变动是一种极为普遍的现象。例如，许多农副产品的产量都因季节更替而有淡季、旺季之分；商业部门许多商品的销售量也随着气候变化的影响而形成了有规则的周期性的变动。季节变动有三个基本特点：一是季节变动每年重复进行；二是季节变动按照一定的周期进行；三是每个周期的变化强度大体相同。

研究季节变动的目的在于了解季节变动对人们经济生活的影响，以便更好地组织生产和安排生活。分析季节变动，还可以根据季节变动的规律，配合适当的季节模型，结合长期趋势进行预测。

一、包含有季节变动的时间数列构成模型

当时间数列中包含长期趋势、季节变动和不规则变动时，其最常用的、基本构成模型的形式是：$Y=T\times S\times I$ 。

确定这一模型时，首先需要配合其长期趋势模型；然后，还需要计算各月(或季度)季节指数；最后将其进行合成，形成时间数列模型。

如果长期趋势为线性趋势模型，季节变动的周期为一年四个季度，则时间数列模型为：

$$Y=T\times S\times I$$

$$\hat{Y}=(a+bt)\times \hat{S}_i,\quad i=1,2,3,4$$

其中，$\hat{S}_i$ 为各月(或季度)季节指数。

所谓季节指数，是某一月份或季度数值与全年平均数值之比。分析季节变动时，主要就是依据这些季节指数，然后根据各季节指数与其平均数(100%)的偏差程度来测定季节变动的程度。如果现象的发展没有季节变动，则各期的季节指数应等于 100%；如果某一月份(或季度)有明显的季节变动，则各期的季节指数应大于或小于 100%。一年中各月(或季度)季节指数之和为 1200%(或 400%)。

二、季节指数的计算

上述模型中，长期趋势模型在前面已有介绍，包括线性趋势模型和非线性趋势模型等。而季节指数的计算最常用的方法主要有简单平均法和趋势剔除法。

(一)简单平均法

简单平均法，是首先要计算各年同期(月或季)发展水平的序时平均数，然后将各年同期平均数与全时期总平均数对比，即得到各期(月或季)的季节指数。

例 14-7：某食品厂近五年各季度的生产量如表 14-10 所示，计算各季度季节指数。

表 14-10　数据资料表　　(单位：万件)

	一季度	二季度	三季度	四季度	合计
2009	10	6	7	9	32
2010	12	8	6	11	37
2011	15	10	8	10	43
2012	13	8	5	12	38
2013	14	9	8	15	46
季度平均数	12.8	8.2	6.8	11.4	9.8
季节指数(%)	130.61	83.67	69.39	116.33	

解：利用简单平均法计算季节指数的方法如下。

首先，各年同一季度的平均数，如一季度为 12.8 万件，二季度为 8.2 万件，三季度为 6.8 万件，四季度为 11.4 万件。

其次，计算各年所有季度总的平均数为 9.8 万件。

第三，计算各季度季节指数：

$$第一季度季节指数=\frac{12.8}{9.8}=130.61\%$$

$$第二季度季节指数=\frac{8.2}{9.8}=83.67\%$$

$$第三季度季节指数=\frac{6.8}{9.8}=69.39\%$$

$$第四季度季节指数=\frac{11.4}{9.8}=116.33\%$$

从时间上看，第一季度和第四季度的季节指数大于 100%，是该产品的生产旺季，而第二季度和第三季度的季节指数小于 100%，是该产品的生产淡季。

简单平均法的优点是计算简便，但其也存在着缺陷。第一，未能消除长期趋势的影响；第二，季节指数的高低受各年数值大小的影响，数值大的年份，对季节指数影响大，数值小的年份，对季节指数的影响小。从上面几个特点看，简单平均法适合于长期趋势是水平趋势的时间数列的季节指数的变动，若时间数列中不仅存在季节变动，同时还存在着上升或下降的长期趋势，用此方法计算的季节指数就会出现偏差。这时可以采用趋势剔除法。

(二) 移动平均趋势剔除法

当时间数列中不仅存在季节变动，同时也存在明显的上升或下降的长期趋势，计算季节指数就需要首先消除长期趋势的影响。剔除长期趋势的方法有很多，如移动平均趋势剔除法、趋势线趋势剔除法等。

移动平均趋势剔除法的基本思想是，先在利用移动平均的方法测定长期趋势后，再将所测定的长期趋势变动从时间数列中予以剔除，并在此基础上计算季节指数。具体的做法是：首先根据各年的月(或季)数据资料计算 12 个月(或 4 个季度)移动平均趋势值 $\hat{T}$，然后将各实际观察值除以相应的趋势值，即 $Y/\hat{T}=S\times I$；再次，将 $S\times I$ 重新按月(或季)排列，求得同月(或季)平均数，即将降低或消除不规则变动，得到各月(或季)季节指数 S。

例 14-8：根据表 14-7 数据，利用移动平均趋势剔除法计算食品生产的季节指数。首先，求出四季度移动平均并进行移正后得到时间数列的趋势值 $\hat{T}$，然后计算 $Y/\hat{T}$。计算过程如表 14-11 所示。

表 14-11　数据资料及计算表　　(单位：万件)

年份	季度	Y	四项移动平均	移正平均(趋势值 T)	Y/T
2011	1	10			
	2	6	8.00		
	3	7	8.50	8.25	84.85
	4	9	9.00	8.75	102.86
2012	1	12	8.75	8.88	135.21
	2	8	9.25	9.00	88.89
	3	6	10.00	9.63	62.34
	4	11	10.50	10.25	107.32

续表

年份	季度	Y	四项移动平均	移正平均(趋势值 T)	Y/T
2013	1	15	11.00	10.75	139.53
	2	10	10.75	10.88	91.95
	3	8	10.25	10.50	76.19
	4	10	9.75	10.00	100.00
2014	1	13	9.00	9.38	138.67
	2	8	9.50	9.25	86.49
	3	5	9.75	9.63	51.95
	4	12	10.00	9.88	121.52
2015	1	14	10.75	10.38	134.94
	2	9	11.50	11.13	80.90
	3	8			
	4	15			

在上面计算的基础上，计算各年同一季度的 Y/T 的平均数，得到季节指数，如表 14-12 所示。

表 14-12　季节指数计算表　（单位：万件）

	一季度	二季度	三季度	四季度
2011			84.85	102.86
2012	135.21	88.89	62.34	107.32
2013	139.53	91.95	76.19	100.00
2014	138.67	86.49	51.95	121.52
2015	134.94	80.90		
平均	137.09	87.06	68.83	107.92
季节指数(%)	136.78	86.86	68.68	107.68

理论上，四个季度的季节指数之和为 400%，本例中四个季度季节指数之和为 400.9%，因此需要校正。校正系数为 400/400.9=0.9977551，用这个系数分别乘表 14-12 中各季度的平均数，即得到表 14-12 中的季节指数。

这种方法由于先消除了长期趋势，所得的季节指数已不受长期趋势的影响，因此测定的季节波动比较精确。

(三)趋势线趋势剔除法

趋势线趋势剔除法的基本思想与移动平均趋势剔除相似，只是在趋势变动测定时采用的不是移动平均的方法，而是趋势线模型的方法。在利用趋势线模型测定长期趋势后，再将所测定的长期趋势变动从时间数列中予以剔除，并在此基础上计算季节指数。具体的做法是：首先根据各年的月(或季)数据资料建立趋势线模型，并计算各月(或季)趋势值 T，然后将各实际观察值除以相应的趋势值，即 $Y/T=S\times I$，再次，将 $S\times I$ 重新按月(或季)排列，求得同月(或季)平均数，即将降低或消除不规则变动，得到各月(或季)季节指数 $\hat{S}$。

利用表 14-10 中的数据资料建立以时间 t 为自变量的趋势方程为：$\hat{T}=8.189+0.153383t$，将时间 t 带入趋势模型中，得到长期趋势测定值 $\hat{T}$，然后计算 $Y/\hat{T}$，计算结果如表 14-13 所示。

表 14-13 数据资料及计算表 （单位：万件）

年份	季度	t	Y	$\hat{T}=8.189+0.153383t$	$Y/\hat{T}$
2011	1	1	10	8.34238	1.198698693
	2	2	6	8.49576	0.706234639
	3	3	7	8.64914	0.80932902
	4	4	9	8.80252	1.022434485
2012	1	5	12	8.9559	1.339898838
	2	6	8	9.10928	0.878225282
	3	7	6	9.26266	0.64776209
	4	8	11	9.41604	1.168219336
2013	1	9	15	9.56942	1.567493119
	2	10	10	9.7228	1.028510306
	3	11	8	9.87618	0.810029789
	4	12	10	10.02956	0.997052712
2014	1	13	13	10.18294	1.276645055
	2	14	8	10.33632	0.773969846
	3	15	5	10.4897	0.476658055
	4	16	12	10.64308	1.127493169
2015	1	17	14	10.79646	1.296721333
	2	18	9	10.94984	0.821929818
	3	19	8	11.10322	0.720511707
	4	20	15	11.2566	1.33255157

在上面计算的基础上，计算各年同一季度的 $Y/\hat{T}$ 的平均数，得到季节指数表，如表 14-14 所示。

表 14-14 季节指数计算表 （单位：万件）

	一季度	二季度	三季度	四季度
2011	1.199	0.706	0.809	1.022
2012	1.340	0.878	0.648	1.168
2013	1.567	1.029	0.810	0.997
2014	1.277	0.774	0.477	1.127
2015	1.297	0.822	0.721	1.333
平均	1.336	0.842	0.693	1.130
季节指数(%)	133.589	84.177	69.286	112.955

第四节 复合型时间数列的分析与预测

进行时间数列分析的目的是，除要了解现象发展变化的规律外，更重要的是，要对未来的发展做预测。当时间数列中既包含有长期趋势，又包含有季节变动和随机变动时，我们可以利用第二节的方法用拟合的长期趋势模型对长期趋势的未来值进行预测；利用第三节的季节指数计算方法计算季节指数；最后利用预测模型 $\hat{Y}=\hat{T}_t\times\hat{S}_i$ 对未来的总体水平做预测。

下面以表 14-10 中的 2011—2015 年各季度的数据资料为例，对 2016—2017 年各季度进行预测。

(1) 首先，利用统计数据建立长期趋势模型如下：

$$\hat{T} = 8.189 + 0.153383t$$

(2) 根据上面长期趋势模型对 2016—2017 年各季度水平进行预测，计算结果如表 14-14 所示。

表 14-14　长期趋势预测值　（单位：万件）

年份	季度	t	$\hat{T} = 8.189 + 0.153383t$
2016	1	21	11.40998
	2	22	11.56336
	3	23	11.71674
	4	24	11.87012
2017	1	25	12.0235
	2	26	12.17688
	3	27	12.33026
	4	28	12.48364

(3) 利用移动趋势剔除法或趋势线趋势剔除法计算各季度的季节指数，表 14-15 所示为利用趋势线趋势剔除法计算的季节指数。

表 14-15　季节指数

	一季度	二季度	三季度	四季度
季节指数(%)	133.589	84.177	69.286	112.955

(4) 利用预测模型 $\hat{Y} = \hat{T}_t \times \hat{S}_i$，预测 2016—2017 年各季度的指标数值，预测结果如表 14-16 所示。

表 14-16　预测结果表　（单位：万件）

年份	季度 i	t	$\hat{T}_t = 8.189 + 0.153383t$	$\hat{s}_i$	$\hat{Y} = T_t \times \hat{S}_i$
	(1)	(2)	(3)	(4)	(5)
2016	1	21	11.40998	133.589	15.24249
	2	22	11.56336	84.177	9.733736
	3	23	11.71674	69.286	8.118039
	4	24	11.87012	112.955	13.4079
2017	1	25	12.0235	133.589	16.06209
	2	26	12.17688	84.177	10.25018
	3	27	12.33026	69.286	8.543121
	4	28	12.48364	112.955	14.1009

思考与练习

1．举例说明时间数列的四个要素及其特点。

2．时间数列分析与回归分析有什么不同？

3．面对一个时间数列，如何将主要构成要素测定出来？

4．利用下面数据资料，绘制时间数列曲线图，配合恰当的曲线趋势方程，对 2012 年的产品产量进行预测。

	2008	2009	2010	2011	2012	2013	2014	2015	2016
产量(吨)	1005	1150	1240	1280	1520	1830	1920	1990	2020

5．某地区 2008—2015 年啤酒产量资料如下：

年份	产量 Y(万吨)
2008	10.54
2009	10.8
2010	10.87
2011	11.16
2012	11.51
2013	12.40
2014	13.61
2015	13.75
合计	94.64

要求：

(1)根据上述资料用最小二乘法拟合该厂啤酒产量的直线趋势方程；

(2)利用所拟合的趋势方程，预测该厂在 2016 和 2017 年啤酒产量的趋势值。

6．已知某化妆品连续三年各月的销售额统计资料如下表所示，要求分析该产品销售额的变化特征，进行时间数列构成要素的分解，并对第四年各月的销售额进行预测。

销售额统计 （单位：万元）

月(第一年)	销售额	月(第二年)	销售额	月(第三年)	销售额
1	9.231	1	12.413	1	14.031
2	7.452	2	11.335	2	10.965
3	6.86	3	11.225	3	10.845
4	5.21	4	7.512	4	6.394
5	3.851	5	5.567	5	4.859
6	3.424	6	3.325	6	4.335
7	2.845	7	2.526	7	3.528
8	1.33	8	4.338	8	6.531
9	12.56	9	13.892	9	12.358
10	16.85	10	15.432	10	15.449
11	17.03	11	16.113	11	16.387
12	11.76	12	12.974	12	15.326

第五部分　统计聚类与数据降维

在统计分析实践中，我们经常会搜集到关于研究对象的众多指标、变量和数据，如何利用这些指标与变量就需要我们进行考虑。分析一个复杂的、综合的问题时，如果只采用其中一个指标，则只能反映研究对象的一个侧面的情况，难免以偏概全；如果要使用多个指标，就要处理这些指标之间的关系，一方面避免由于简单的综合而导致信息的重叠，另一方面要将复杂问题简单化。以下两章的内容主要介绍如何利用多个指标对研究对象进行聚类，如何将众多未经系统化梳理的指标进行梳理、降维，以及如何将复杂问题简单化的基本方法。

第十五章　聚类分析和判别分析

第一节　统计聚类分析

一、聚类的基本思想

(一)聚类分析

物以类聚，人以群分。在现实的社会、经济、管理分析中，存在着大量的分类需求。例如，为了研究消费者的行为，需要对消费者进行分类；为了进行产品的研究，需要对产品进行分类；为了进行更准确的市场细分，需要对各地市场进行分类等。在传统的分类中，有根据单一指标的简单分组，也有根据多指标进行的复合分组。对复杂的、多指标的研究对象进行分类时，主要是根据经验和专业知识进行主观判断分类。但是，靠经验和专业知识进行分类处理，会使许多分类带有主观性和任意性。随着社会经济的发展及科学的进步，我们掌握的信息越来越多，进行数据处理的方法与手段也越来越先进，同时对分类的需求也越来越高。数理统计的多元分析方法也有了迅速的发展，并被引入分类方法中，因而逐渐形成了多指标的聚类方法。

所谓聚类分析，又称为群分析和分类分析，是一种重要的分类方法。它是根据事物的特征及各事物之间的近似性，在多指标、多信息源的基础上，通过已建立的统计模型对事物进行归类的一种多元统计分析方法。

(二)聚类的基本思想

聚类分析是一种定量的分析方法。对研究对象各单位聚类前，我们应先确定反映其基本特征的可计量的指标。如果研究对象中各单位存在多种可计量的指标反映其特征，我们就可依据反映这些特征的多个指标按近似性将其进行归类。为了说明聚类的基本思想，下面我们把问题简化，看一个最简单的、直观的、容易理解的例子。

现已知有某集团下属甲、乙、丙、丁四个分公司的年收入额资料如表 15-1 所示。

表 15-1　各分公司年收入额

公司	年收入(百万元)
甲	1
乙	3
丙	9
丁	14

若我们仅根据年收入额资料进行分类，则可把它们排列在数轴上，每个分公司作为一类，即类(甲)、类(乙)、类(丙)、类(丁)，情况如图 15-1 所示。

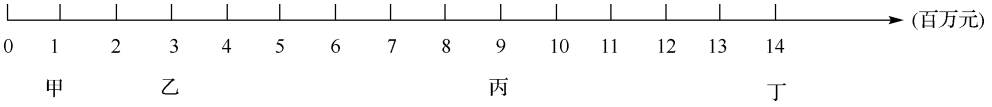


图 15-1　根据年收入额分类

若根据距离的远近(数据差异大小)进行聚类，我们从图 15-1 中可以直观地观察到甲和乙的距离最近，它们的距离是 3−1=2。这样，聚类时首先就要把甲和乙两个分公司聚合成一个类，以后为了方便，称之为“类(甲乙)”，这样就形成了三类，即：类(甲乙)、类(丙)、类(丁)。

我们把甲、乙聚为一类的同时，在图 15-1 中再增加一个表示距离的维度，如图 15-2 所示。

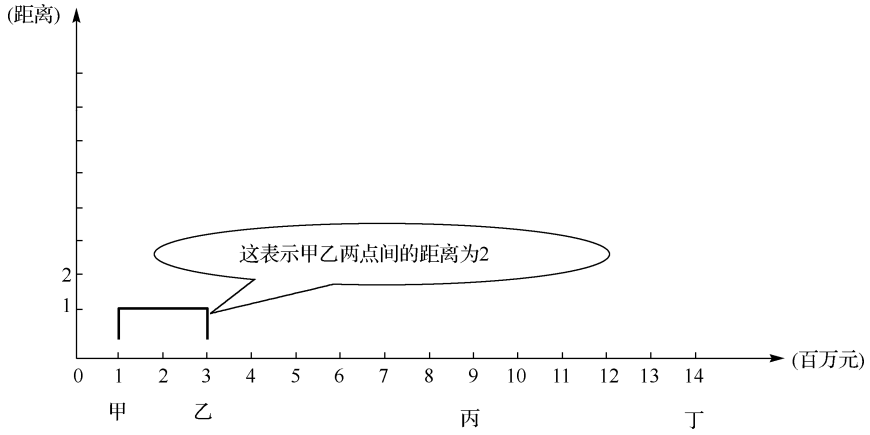


图 15-2　增加距离维度的分类

接下来，我们继续观察，发现剩下的点里，丙、丁的距离最近，其距离为 14–9=5。因此聚为“类(丙丁)”。这样，初始的四类就形成两类，即类(甲乙)、类(丙丁)，如图 15-3 所示。

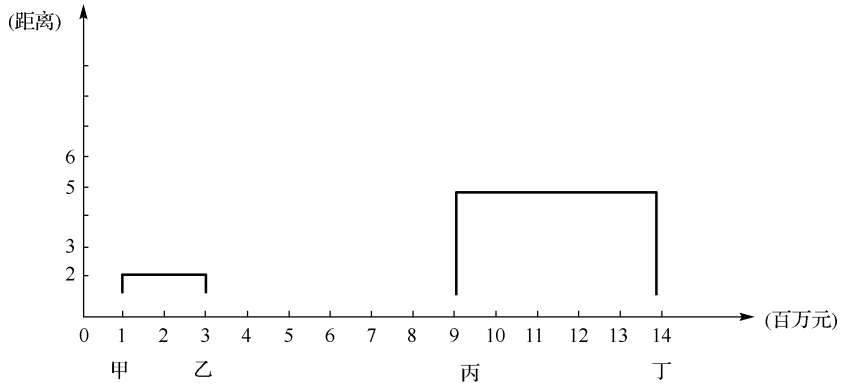


图 15-3　进一步聚类

如果我们按照最短距离的原则(即将两类别中两两单位之间距离最短的距离，作为这两类之间的距离，详细解释见后面的系统聚类法)来归类，那么乙、丙之间的距离，就作为“类(甲乙)”和“类(丙丁)”的距离，再把这两类聚在一起，就成为一类。这样整个聚类过程就完成了，其结果如图 15-4 所示。这种聚类的方法属于系统聚类。

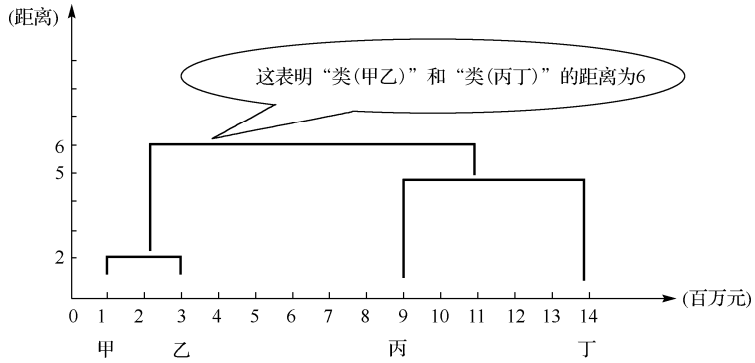


图 15-4 聚类结果

上面就是一个最简单的、最直观的一个聚类过程示意性的例子。

虽然这是一个非常简单的示意性的例子，但从中可以看出聚类分析就是要将距离最近(或相近程度最高)的单位(点或类)聚为一类。当然在实际进行聚类分析时，往往要考虑多个指标(变量)。而在根据多个指标进行聚类时，衡量这个“距离远近”或“相近程度”，就是要根据“距离”或“相似系数”来确定。

聚类分析虽然比传统的分类方法有明显的优势，但仍是一种探索性的分析方法。其在理论上与一般的推断统计相比并不完善，在方法上也有人认为比较粗糙，但由于其可以解决许多实际问题，所以受到普遍的重视与应用。聚类分析不仅可以用来对各单位、样品(个体)进行分类，也可以用来对变量进行分类，对单位、样品(个体)的分类一般称为“Q 型聚类”，而对变量的分类称为“R 型聚类”。

二、距离与相似性度量

对研究对象中各单位进行聚类分析，需要了解用于聚类的指标数据的类型，不同类型的指标数据在聚类分析中处理的方式不同。通常，度量指标数据的尺度有三种类型：(1)间隔尺度(第二章提及的定距和定比尺度)，即指标值表现为数值型数据，如收入、利润、产值等；(2)定序尺度，指标值表现为顺序数据，如文化程度、满意度等；(3)名义尺度(即定类尺度)，指标值表现为分类数据，如性别、籍贯等。总的来说，现有的度量数值型数据的方法比较多，而分类数据、顺序数据的处理方式相对比较少。本节介绍最常见的数据类型在聚类分析时采用的是近似程度指标。

近似系数包括衡量各样本单位之间“远近”的距离和衡量变量之间“相似程度”的相似系数两类。

(一)样本单元之间的近似性度量

样本单元之间的近似度，是指以研究对象中的样本单元为单位，计算两两样本单元之间的近似程度。

1. 距离

聚类分析时，如果每个样本单元有 p 个指标，则每个样本单元 x_i 可以看成 p 维空间中的一个点，n 个样本单元就组成 p 维空间中的 n 个点。用 x_{ij} 表示第 i 个样本单元的第 j 个指标，用 $\bar{x}_j$ 表示第 j 个指标的均值。常见的测量样本单元之间距离的方法如下。

(1)明考斯基距离(明氏距离)。

明氏距离(Minkowski)是最常用的距离之一，其计算公式为：

$$d_{ij}(q)=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{q}\right)^{\frac{1}{q}}$$

当 q=1 时，$d_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$，称为绝对距离（Absolute Distance）；

当 q=2 时，$d_{ij}(2)=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^{2}\right)^{\frac{1}{2}}$，称为欧氏距离（Euclidean Distance）；

当 q=∞时，$d_{ij}(\infty)=\max_{1\leqslant k\leqslant p}\left|x_{ik}-x_{kj}\right|$，称为切比雪夫距离（Chebyshev Distance）。

明氏距离的优点是简单且易于理解，但计算结果受各聚类指标计量单位的影响较大。例如，要根据长度和重量两个指标计算两样本单元之间的明氏距离，在采用不同的计量单位时，其距离测量的结果不同。已知样品 A、B、C、D 的评价得分和长度（cm）指标的测量结果分别是 A（0，10）、B（1，0）、C（0，5）、D（10，0），如图 15-5 所示。

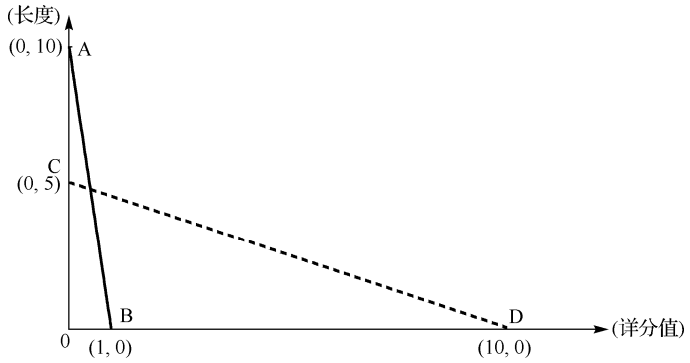


图 15-5　距离测量结果

以欧氏距离为例：

$$d_{12}(2)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2}$$

当长度用 cm 度量时，AB、CD 之间的距离分别为：

$$d_{\mathrm{AB}}=\sqrt{(0-10)^2+(1-0)^2}=\sqrt{101}$$

$$d_{\mathrm{CD}}=\sqrt{(0-5)^2+(10-0)^2}=\sqrt{125}$$

所以　　$d_{\mathrm{AB}}<d_{\mathrm{CD}}$

当长度用 mm 度量时，AB、CD 之间的距离分别为：

$$d_{\mathrm{AB}}=\sqrt{(0-100)^2+(1-0)^2}=\sqrt{10001}$$

$$d_{\mathrm{CD}}=\sqrt{(0-50)^2+(10-0)^2}=\sqrt{2600}$$

所以　　$d_{\mathrm{AB}}>d_{\mathrm{CD}}$

由此可见，明氏距离的大小受聚类指标计量单位的影响。为避免由于计量单位或指标的量纲对计算结果产生影响，实际应用时，可以先对数据进行标准化处理。

用 $\bar{x}_j$ 表示第 j 个指标的均值，用 S_j 表示第 j 个指标的标准差，则标准化后的数据为：

$$x_{ij}^{/}=\frac{x_{ij}-\overline{x}_j}{S_j}$$

然后用标准化后的数据计算距离。

明氏距离的另一个缺点是它没有考虑到指标之间的相关性。

(2)马氏距离。

马氏距离是 1936 年印度数学家马哈拉比斯由协方差矩阵构造并计算的距离，它的设计不仅考虑到了不同指标间量纲不同的问题，也解决了明氏距离没有考虑到的指标之间的相关性的缺陷。其计算公式为：

$$d_{ij}(M)=(\boldsymbol{X}_i-\boldsymbol{X}_j)'\Sigma^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j)$$

式中，$\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 分别为 i 和 j 两个样本单元指标的矩阵行向量的转置；Σ^{-1} 为数据矩阵的协方差阵。

马氏距离对一切线性变换是不变的，故它不受指标量纲的影响，同时它对指标的相关性也做了考虑。

上面的距离是适用于数值型变量的，如果是顺序数据或分类数据，也有一些定义距离的方法，可参考多元统计分析等方面的教材，本节略。

2. 相似系数

在聚类分析中不仅需要将样本单元分类，有时候也需要对指标分类，在指标之间也可以定义距离。但更常用的是相似系数，性质越接近的样本单元或指标，它们的相似系数的绝对值越接近 1，而彼此无关的样品或指标，它们的相似系数的绝对值越接近于零。

聚类时常用的相似系数有夹角余弦系数和相关系数。

(1)夹角余弦系数。

尽管图 15-6 中 *AB* 和 *CD* 长度不一样，但形状相似。当长度不是主要矛盾时，就可利用夹角余弦这样的相似系数。

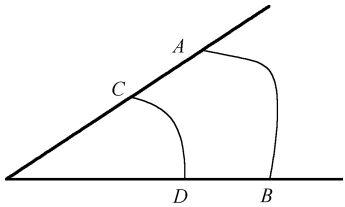


图 15-6　夹角余弦系数

夹角余弦的计算公式为：

$$\cos\theta_{ij}=\frac{\sum_{k=1}^{p}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{p}x_{ik}^2\sum_{k=1}^{p}x_{jk}^2}}$$

当 $\cos\theta_{ij}=1$，说明两个样品 x_i 与 x_j 完全相似；$\cos\theta_{ij}$ 接近 1，说明 x_i 与 x_j 相似程度高；$\cos\theta_{ij}=0$，说明 x_i 与 x_j 完全不一样；$\cos\theta_{ij}$ 接近 0，说明 x_i 与 x_j 差别大。

(2)相关系数。

所谓相关系数(在相关与回归分析一章中有详细介绍)是指测量两变量之间相关密切程度的

一个统计量。聚类分析应用中我们常用相关系数来刻画指标之间的相似性。聚类时，首先把指标之间的相关系数都计算出来，然后将相关系数最大的两类合并成一类，并以此方式逐步归类。

相关系数的计算公式为：

$$r_{ij}=\frac{\sum_{k=1}^{p}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j)}{\sqrt{\sum_{k=1}^{p}(x_{ik}-\overline{x}_i)^2\sum_{k=1}^{p}(x_{jk}-\overline{x}_j)^2}}$$

当 r_{ij}=1，说明两个样品 x_i 与 x_j 完全相似； r_{ij} 接近 1，说明 x_i 与 x_j 相似程度高； r_{ij}=0，说明 x_i 与 x_j 完全不相关； r_{ij} 接近 0，说明 x_i 与 x_j 差别大。

(二)类之间的近似性度量

所谓类，是相似的样本单元(或指标)的集合，在聚类分析中，不仅要计算样本单元之间的距离或相似系数，也需要计算类之间的距离或相似系数。

若用 G 表示类，其中包含若干个样品，分别用 $x_1,x_2\cdots$ 表示， D_{pq} 表示第 p 类 G_p 和第 q 类 G_q 之间的距离，度量类间距离的方式常用的有：

1. 最短距离

$$D_{pq}=\min d(x_i,x_j)$$

两类之间的距离是类 G_p 和类 G_q 中最临近的两个样本单元间的距离，可以用图 15-7 表示。

图 15-7　两类之间的距离

例 15-1：设有五个样品，每个样品的数据分别是 1、2、3.5、7、9。现用最短距离法对五个样品进行分类。

首先，采用绝对距离的方法，计算样品两两之间的距离，得到如下距离阵(如图 15-8 所示)。

	G1	G2	G3	G4	G5
G1	0				
G2	1	0			
G3	2.5	1.5	0		
G4	6	5	3.5	0	
G5	8	7	5.5	2	0

图 15-8　距离阵

从计算结果可知，G1 和 G2 的距离最近，距离为 1，因此将 G1 和 G2 合并成新的一类，记为 G6。接下来如果将 G6 与其他类再进行合并呢？最短距离法就是在这两类中所有对应其他类的距离中取最小的数值。如 G1 与 G3 的距离是 2.5，G2 与 G3 的距离是 1.5，取它们俩之中最小的数值，即 1.5 作为新类 G6 与 G3 的距离，其余类推(如图 15-9 所示)，然后将最近的两类合并。

	G6	G3	G4	G5
G6	0			
G3	1.5	0		
G4	5	3.5	0	
G5	7	5.5	2	0

图 15-9　按距离分类结果

2. 最长距离

两类之间的距离是类 G_p 和类 G_q 中最临近的两个样品的距离，可以用图 15-10 表示。

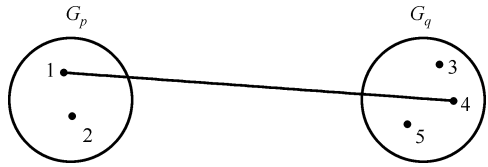


图 15-10　最长距离

还以上例来说明最长距离法的思路。

首先，还是采用绝对距离的方法，计算样本单元两两之间的距离，得到如图 15-11 所示的距离阵。

	G6	G3	G4	G5
G6	0			
G3	2.5	0		
G4	6	3.5	0	
G5	8	5. 5	2	0

图 15-11　距离阵

然后进行类与类的归类。从计算结果可知，G1 和 G2 的距离最近，距离为 1，因此将 G1 和 G2 合并成新的一类，记为 G6。最长距离法就是在这两类中所有对应其他类的距离中取最大的数值。如 G1 与 G3 的距离是 2.5，G2 与 G3 的距离是 1.5，取它们两个之中最大的数值，即 2.5 作为新类 G6 与 G3 的距离，其余类推(如图 15-11 所示)。

3. 类平均法

类平均法是采用类 G_p 和类 G_q 中任两个样本单元距离的平均，图 15-12 中两类之间的距离为：

$$D_{pq}=\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$$

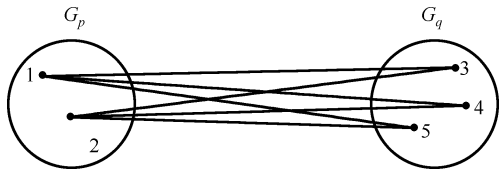


图 15-12　G_p、G_q 两类之间的距离

除此以外，类间距离的测量还有其他方法，例如，重心法就是在合并新类时计算要合并类之间距离的平均值；离差平方和法(也称 Ward's Method)的思想来源于方差分析，即同类样本单元的离差平方和等。

三、聚类的基本方法

在样品(或指标)距离和相似系数度量的基础上进行聚类，有多种可选择方法，如系统聚类、模糊聚类、*k* 均值聚类等，本节下面将介绍最常用的两种方法。

(一) 系统聚类法

系统聚类法的基本思路是，先将研究对象中每个样本单元或指标各自看成一类，共 *n* 类。然后根据样本单元间的相似度量，将 *n* 类中最相似的两类合并，组成一个新类，这样得到 *n*–1 类，再在这 *n*–1 类中找出最相似的两类合并，得到 *n*–2 类，以此类推，直至将所有的样本单元合并成一个大类为止。

系统聚类的步骤可以用下面的流程图(如图 15-13 所示)来表示。

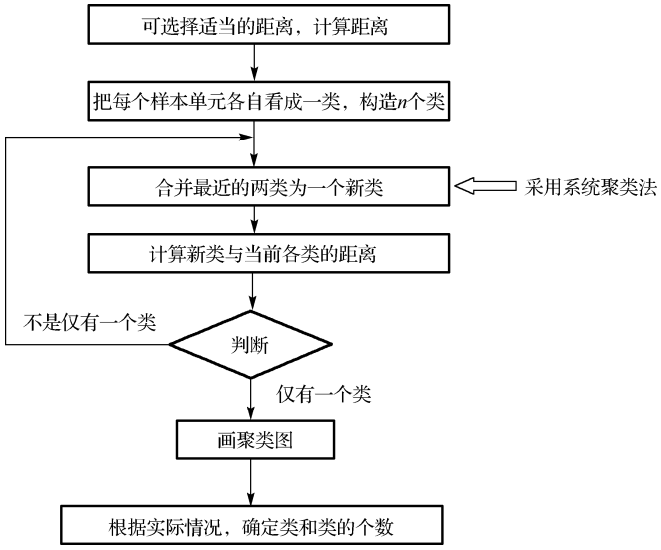


图 15-13　系统聚类的步骤

下面按上面的流程，根据 20 名学生的数学、物理、化学、语文、历史和英语成绩对其进行聚类。已知 20 名学生的成绩数据资料如图 15-14 所示。

	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
10	86	94	97	51	63	55
11	74	80	88	64	73	66
12	67	84	53	58	66	56
13	81	62	69	56	66	52
14	71	64	94	52	61	52
15	78	96	81	80	89	76
16	69	56	67	75	94	80
17	77	90	80	68	66	60
18	84	67	75	60	70	63
19	62	67	83	71	85	77
20	74	65	75	72	90	73

图 15-14　成绩截图

利用 SPSS 软件中的系统聚类法完成基本的聚类分析计算过程。

分层聚类法既可对样本单元(Cases)进行聚类，也可对变量(Variables)聚类。这两种聚类的方法和原理是一致的。具体的操作如下。

SPSS 选项：【Analyze】—【Classify】—【Hierarchical Cluster】。

第一步：将所有六门课程的成绩变量放入【Variables】中，并在样本单元聚类和变量聚类中选择样本单元聚类。

第二步：单击【Method】按钮进入，在【Cluster Method】中选择【Ward's Method】或其他方法，并选择数据的标准化的方法。

第三步：单击【Plot】按钮进入，选择【Dendrogram】。

第四步：单击【OK】按钮即可。

SPSS 输出结果如图 15-15(谱系图)所示。

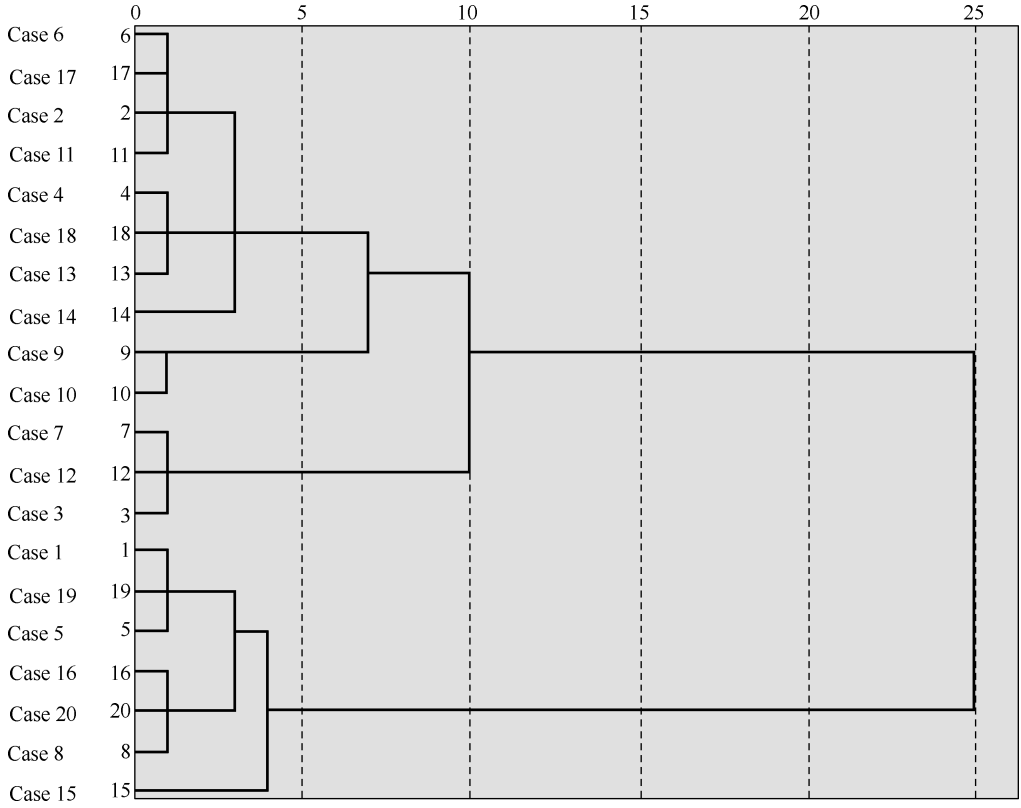


图 15-15 聚类分析的谱系图

从谱系图看所有的样本单元(学生)最后聚成一大类，我们可用一支笔从上到下去竖切这个图，这支笔切过几条线，就表示分几类。例如，我们在标尺的 15 处从上到下竖切时，就会切过两条线，表示可将原样本单元分成两大类；又如，我们在标尺的 8 附近从上到下竖切时，就会穿过三条线，即可将原样本单元分成三大类。那么到底将这些样本单元分两类好还是分三类好？只从聚类图看是不会有结论的。因此我们必须将分成两类和三类的结果存入数据文件中，然后进行每一类的均值分析。

SPSS 的实现：

SPSS 选项：【Analyze】—【Classify】—【Hierarchical Cluster】。

第一步：单击【Save】按钮进入，在【Cluster Membership】中选择【Rang of Solution】，在【From】后输入 2，在【Through】后输入 3(表示存入从两类到三类的分类结果)。

第二步：单击【OK】按钮即可。

SPSS 选项：【Analyze】—【Compare Means】—【Means】。

第一步：将刚存入的分类变量 clu2_1、clu3_1 放入【Independent List】中。

第二步：将所有态度变量放入【Dependent List】中。

第三步：单击【OK】按钮即可。

将 20 名学生分三类和分两类的输出结果如表 15-2 所示。

表 15-2　聚类成员

个案	分三聚类	分两聚类
1：Case 1	1	1
2：Case 2	2	2
3：Case 3	3	2
4：Case 4	2	2
5：Case 5	1	1
6：Case 6	2	2
7：Case 7	3	2
8：Case 8	1	1
9：Case 9	2	2
10：Case 10	2	2
11：Case 11	2	2
12：Case 12	3	2
13：Case 13	2	2
14：Case 14	2	2
15：Case 15	1	1
16：Case 16	1	1
17：Case 17	2	2
18：Case 18	2	2
19：Case 19	1	1
20：Case 20	1	1

各类指标的均值及均值差异的显著性分析(方差分析)结果如表 15-3、表 15-4、表 15-5、表 15-6 所示。

表 15-3　样本单元(学生)分三类时各类的均值

		数学	物理	化学	语文	历史	英语
1	平均值	71.29	69.43	73.57	76.86	86.57	75.71
	个案数	7	7	7	7	7	7
	标准差	6.102	12.817	9.199	5.728	4.791	3.251

续表

		数学	物理	化学	语文	历史	英语
2	平均值	74.73	73.73	72.82	61.36	68.09	58.64
	个案数	11	11	11	11	11	11
	标准差	6.198	9.737	13.280	6.697	3.859	4.864
3	平均值	84.50	97.00	88.00	46.00	65.00	52.50
	个案数	2	2	2	2	2	2
	标准差	2.121	4.243	12.728	7.071	2.828	3.536
总计	平均值	74.50	74.55	74.60	65.25	74.25	64.00
	个案数	20	20	20	20	20	20
	标准差	6.825	12.878	12.215	11.575	10.120	9.878

表 15-4 将样本单元(学生)分为三类的类均值差异分析表

		平方和	自由度	均方	*F*	显著性
数学	组间	468.005	2	234.002	9.540	.002
	组内	416.995	17	24.529		
	总计	885.000	19			
物理	组间	366.469	2	183.235	1.119	.350
	组内	2784.481	17	163.793		
	总计	3150.950	19			
化学	组间	1392.819	2	696.410	8.210	.003
	组内	1441.981	17	84.822		
	总计	2834.800	19			
语文	组间	1452.626	2	726.313	11.295	.001
	组内	1093.124	17	64.301		
	总计	2545.750	19			
历史	组间	1645.136	2	822.568	46.517	.000
	组内	300.614	17	17.683		
	总计	1945.750	19			
英语	组间	1481.905	2	740.952	33.852	.000
	组内	372.095	17	21.888		
	总计	1854.000	19			

表 15-5 将样本单元(学生)分两类时各类的均值

		数学	物理	化学	语文	历史	英语
1	平均值	71.29	69.43	73.57	76.86	86.57	75.71
	个案数	7	7	7	7	7	7
	标准差	6.102	12.817	9.199	5.728	4.791	3.251
2	平均值	76.23	77.31	75.15	59.00	67.62	57.69
	个案数	13	13	13	13	13	13
	标准差	6.772	12.526	13.892	8.651	3.798	5.105
总计	平均值	74.50	74.55	74.60	65.25	74.25	64.00
	个案数	20	20	20	20	20	20
	标准差	6.825	12.878	12.215	11.575	10.120	9.878

表 15-6　将样本单元(学生)为两类的类均值差异分析表

		平方和	自由度	均方	*F*	显著性
数学	组间	111.264	1	111.264	2.588	.125
	组内	773.736	18	42.985		
	总计	885.000	19			
物理	组间	282.466	1	282.466	1.773	.200
	组内	2868.484	18	159.360		
	总计	3150.950	19			
化学	组间	11.393	1	11.393	.073	.791
	组内	2823.407	18	156.856		
	总计	2834.800	19			
语文	组间	1450.893	1	1450.893	23.853	.000
	组内	1094.857	18	60.825		
	总计	2545.750	19			
历史	组间	1634.959	1	1634.959	94.691	.000
	组内	310.791	18	17.266		
	总计	1945.750	19			
英语	组间	1477.802	1	1477.802	70.709	.000
	组内	376.198	18	20.900		
	总计	1854.000	19			

分两类还是分三类合理呢？比较两种分类结果(分两类和分三类)的方差分析结果(具体分析原理和分析方法参看方差分析一章)，我们可以看到分三类时，不同类别之间六门课程的差异均是显著的，但是分两类的话，数学、物理、化学三门课程在不同类别之间的差异并不显著。故按六门课成绩分类，将学生分成三类是比较恰当的。

(二) *K*-均值聚类法

K-均值聚类法，也称快速聚类法。快速聚类法的过程与分层聚类法不同，它在聚类前要先确定分几类，然后利用迭代的方法进行聚类。具体过程如下。

第一步，选择 n 个数值型变量参与聚类分析，确定要求的聚类数为 k 个。

第二步，由系统选择 k 个(聚类的类数)观测量(也可由用户指定)作为聚类的种子。

第三步，按照距这些类中心距离最小的原则，把所有样本单元分派到各类中心所在的类中。

第四步，这样每类中可能会分配到若干个样本单元，再计算每个类中各个变量的均值，以此作为第二次迭代的中心。

第五步，然后根据这个中心重复第三步、第四步，直到中心的迭代标准达到要求时(一般是当所有的类中心不再变化时)，则聚类过程结束。

现在我们仍以学生成绩为例，来说明快速聚类的过程。在上例中我们已知将学生分为三类，因此在快速聚类时，确定聚类的个数 k=3。

SPSS 的实现：

SPSS 选项：【Analyze】—【Classify】—【K-Means Cluster】。

第一步：在【Number of Cluster】中填入 3。

第二步：将所有的态度变量放入【Variables】中。

第三步：单击【Save】按钮进入，选择【Cluster Menbership】。
第四步：单击【OK】按钮即可。

SPSS 输出结果如下：表 15-7(1)是初始类中心；表 15-7(2)是类中心的迭代过程，从表中可知，只迭代了两次；表 15-7(3)是最终类的中心，比较表 15-7(1)和表 15-7(3)的类中心确实发生了变化。

表 15-7(1) 初始聚类中心

	聚类		
	1	2	3
数学	69	86	67
物理	56	94	84
化学	67	97	53
语文	75	51	58
历史	94	63	66
英语	80	55	56

表 15-7(2) 迭代历史记录①

迭代	聚类中心中的变动		
	1	2	3
1	18.805	15.778	19.211
2	.000	.000	.000

① 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 2。初始中心之间的最小距离为 49.558。

表 15-7(3) 最终聚类中心

	1	2	3
数学	72	78	74
物理	69	86	73
化学	74	88	66
语文	77	55	60
历史	85	66	68
英语	74	57	58

虽然都是利用多指标进行聚类，但分层聚类法与快速聚类法在应用上是存在着一定区别的。

层次聚类法的聚类过程是单方向的，一旦某个样品(Case)进入某一类，就不可能从该类出来，再归入其他的类。

而快速聚类法受奇异值、相似测度等的影响使得聚类变量的影响较小，对于不合适的初始分类可以进行反复调整。

在聚类分析发展的早期，层次聚类法应用普遍，其中尤其以组间类平均法与离差平方和法应用最广。

后来快速聚类方法逐步被人们接受，应用日益增多。现在是层次聚类法和快速聚类法两者相结合，取长补短。例如，首先使用层次聚类法确定分类数，检查是否有奇异值，去除奇异值后，对剩下的样本单位重新进行分类，然后再应用快速聚类法对样本进行重新调整。

第二节　判 别 分 析

判别分析和前面的聚类分析的不同主要表现为：在聚类分析中，一般人们事先并不知道或明确一定要分成几类，完全根据数据来确定。

而在判别分析中，已经有了一个明确知道类别的“样本”，利用这个样本数据，就可以建立判别准则，并对未知类别的观测值进行判别。

判别分析被广泛地应用于各领域中。例如，在医学上用计算机看病，就是将各种普通病症的特征输入到计算机中，建立判别准则，然后对病人进行初步诊断。又如在税收上，判别分析可以帮助鉴别税收交纳的情况。即可以将不同经营行业、规模、地点的企业的经营状况数据和应纳税情况作为已知类别的样本数据，然后再建立判别准则，对其他纳税的企业进行判别。如果出现严重不符，就可能存在着偷税或漏税的问题，需要进一步审查。

一、判别分析的基本思路

设有 $G_1,G_2,\cdots G_K$ 个总体，从不同的总体中抽出不同的样本，根据样本数据建立判别法则，然后利用该法则判别新的样品属于哪一个总体。当然，根据不同的方法，建立的判别法则也是不同的。常用的判别方法有距离判别法和 Fisher 判别法。

(一)距离判别法(不用投影)

由于已经知道样本数据的类别了，所以可以求得每个类别的中心。这样只要定义了如何计算距离，就可以得到任何给定的样本数据到各类中心的距离。

显然，最简单的办法就是离哪个类的中心距离最近，就属于哪一类。通常使用的距离是马氏(Mahalanobis)距离。

用来比较到各个中心距离的数学函数称为判别函数(Discriminant Function)。这种根据远近判别的方法，原理简单，直观易懂。

假设有两个总体 G_1 和 G_2，如果能够定义点 x 到它们的距离 $D(x,G_1)$ 和 $D(x,G_2)$，则：

如果 $D(x,G_1) < D(x,G_2)$，则 $x \in G_1$；

如果 $D(x,G_2) < D(x,G_1)$，则 $x \in G_2$；

如果 $D(x,G_1) = D(x,G_2)$，则待判。

(二)Fisher 判别法(先进行投影)

所谓 Fisher 判别法，就是一种先投影的方法。考虑只有两个(预测)变量的判别分析问题。假定这里只有两类。数据中的每个观测值都是二维空间的一个点，如图 15-16 所示。

这里只有两种已知类型的样本数据。其中一类有 38 个点(用“o”表示)，另一类有 44 个点(用“*”表示)。按照原来的变量(横坐标和纵坐标)，很难将这两种点分开。

于是就寻找一个方向，也就是图上的虚线方向，沿着这个方向朝和这个虚线垂直的一条直线进行投影会使得这两类分得最清楚。可以看出，如果向其他方向投影，判别效果不会比这个好。

有了投影之后，再用前面讲到的距离远近的方法来得到判别准则。这种首先进行投影的判别方法就是 Fisher 判别法。

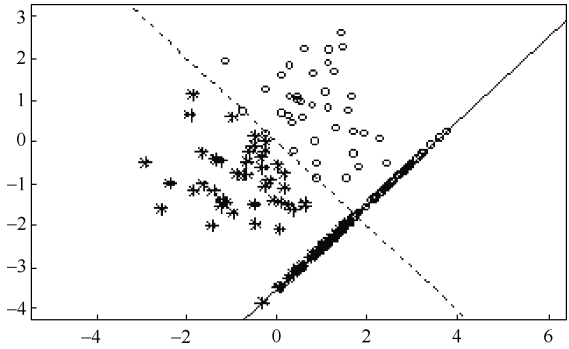


图 15-16 Fisher 判别法

二、判别分析的基本模型

判别分析的基本模型是判别函数：

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

式中，y 为判别值；x_i 为各判别变量；b_i 为相应的判别系数，它表示各判别变量对于判别函数值的影响。

在判别分析时，判别变量较多时，判别函数也往往有多个。

(一)模型估计过程的简略描述

首先将判别变量进行旋转，寻找某个角度使各类的平均值的差别尽可能大，然后将其作为判别的第一维度(即投影)。在这一维度上可以代表或解释原始变量组间方差中最大的部分。对应第一维度的判别函数称为第一判别函数。

然后按照同一原则寻找第二维度，并建立第二判别函数。

如此下去，直至推导出所有的判别函数。

(二)判别分析的假设条件

(1)样本数据的分组类型在两种及以上，即 $g \geqslant 2$；

(2)各判别变量必须是定量变量，并且要求观察值数量比变量的个数至少多两个(cases≥k+2)；

(3)每一个判别变量不能是其他判别变量的线性组合；

(4)各组总体的协方差阵相等；

(5)各判别变量之间具有多元正态分布。

三、判别分析的 SPSS 实现

下面举例说明判别分析 SPSS 的实现。

例 15-2：通常对工业企业的经济效益进行考核和评价的指标有：总资产贡献率、资产负债率、流动资产周转率和工业成本费用利润率等。根据经济效益的好坏可以分成三类：效益好的企业、效益一般的企业和效益较差的企业。现在我们分别搜集到这三类企业的经济效益数据。利用这三种类型的企业样本数据建立判别函数，然后对新样品数据进行判别。其中已分类的企业(样品数据)为 60 个，有两个企业(样品数据)为待判数据。

SPSS 的实现：

SPSS 选项：【Analyze】—【Classify】—【Discriminant】。

第一步：将变量“类型”放入【Grouping Variable】中，并单击【Define Range】定义取值范围，最小值 1，最大值 3。

第二步：将反映经济效益的四个变量放入【Independent】中。

第三步：单击【Statistic】按钮进入，在【Descriptive】中选择：【Univariate ANOVAs】(均值检验)；在【Function Coefficients】中选择【Fishers】(Fishers 系数)和【Unstandardized】(非标准的判别系数)。

第四步：单击【Classify】按钮进入，在【Display】中选择【Summary Table】(分类结果)和【Leave-one-out classification】；在【Plots】中选择【Combined-Groups】。

第五步：单击【Save】按钮进入，选择【Predictor Group Membership】。

第六步：单击【OK】按钮即可。

SPSS 输出结果如下。

(1) 表 15-8(1) 输出的是各组均值相等的检验。这里是应用方差分析进行均值检验。从检验结果看，所有变量的 Sig=0.000<0.05，因此说明反映经济效益的这四个变量在不同类别中有显著的差异。

表 15-8(1)　组均值差异的检验

	Wilks' Lambda 值	*F*	df1	df2	Sig.
总资产贡献率	0.218	102.321	2	57	0.000
资产负债率	0.503	28.166	2	57	0.000
流动资产周转率	0.246	87.310	2	57	0.000
工业成本费用利润率	0.079	331.401	2	57	0.000

(2) 表 15-8(2) 是协方差阵相等的检验。判别分析应用的前提条件之一就是各组协方差阵应该相等。但是检验结果表明，各组的协方差阵差异显著，没有符合判别分析的协方差阵相等的前提条件。但在实际中这种前提条件很难满足，所以我们仍可进行判别分析，最后可由判别的效果来决定是否采用该判别分析。

表 15-8(2)　检验结果

Box's M(等方差检验)		86.749
F	Approx.(近似值)	3.894
	df1	20
	df2	11662.473
	Sig.	0.000

注：Tests null hypothesis of equal population covariance matrices(原假设为总体协方差阵相等)。

(3) 表 15-8(3) 是判别力指数。本例有两个判别函数用于分析。每个判别函数的判别能力大小是由其特征值的大小决定的，该值越大说明判别函数包含原始样本数据中的差异越大，其判别能力越强。此外，从每个判别函数的方差贡献率来看，第一个判别函数的判别信息已占所有判别信息的 84.3%，而第二个判别函数只占 15.7%。

(4) 表 15-8(4) 输出的是残余判别力。本表说明建立某判别函数之前的数据中所剩余的

判别信息，λ(Lambda)值越小，说明数据中所剩余的判别信息越多。表中第一行输出结果表明：在建立判别函数之前，原始信息中所剩余的判别信息 λ 值等于 0.017，非常小，说明数据中的判别信息较强；表中第二行输出结果表明在建立第一个判别函数以后，即建立第二个判别函数之前，数据中所剩余的判别信息 λ 值等于 0.266，明显增大。但是从检验结果来看，第二个判别函数的卡方检验仍然显著，因而这两个判别函数都有效。

表 15-8(3) 特征值

Function	Eigenvalue 特征值	% of Variance 方差贡献率	Cumulative % 累计方差贡献率	Canonical Correlation
1	14.809(a)	84.3	84.3	0.968
2	2.753(a)	15.7	100.0	0.856

a First 2 canonical discriminant functions were used in the analysis(分析中用到了 2 个典型判别函数)。

表 15-8(4) Wilks' Lambda 值

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	0.017	226.616	8	0.000
2	0.266	73.403	3	0.000

(5)表 15-8(5)是非标准化判别函数系数表。通过它可以计算各判别值，并且可用来作图，表示在判别空间中各样品点的位置。

表 15-8(5) Canonical Discriminant Function Coefficients(典型判别函数系数)

	Function	
	1	2
总资产贡献率	0.260	0.010
资产负债率	0.040	0.002
流动资产周转率	−1.805	4.915
工业成本费用利润率	0.725	0.146
(Constant)	−8.008	−9.853

Unstandardized coefficients(非标准化系数)。

(6)图 15-17 是判别结果图形。

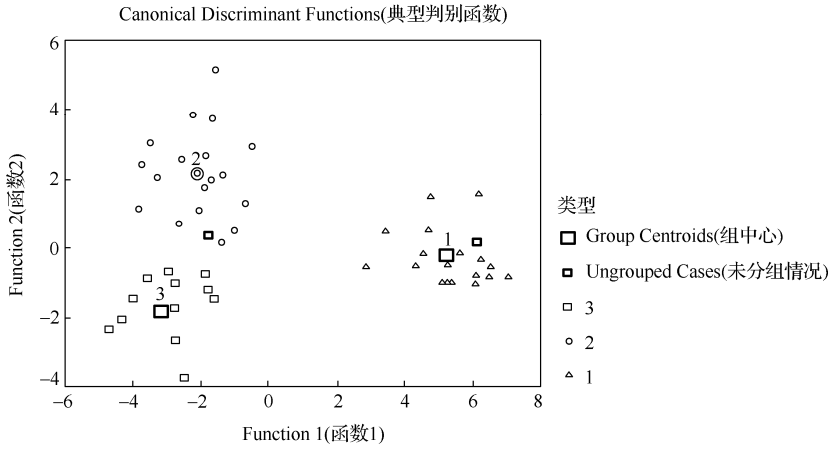


图 15-17 判别结果图形

从图形中可直观地看到各种类型的判别结果分布情况。图中的 Function1 表示第一个判别函数，Function2 表示第二个判别函数。

(7)表 15-8(6)是标准化判别函数系数表。通过标准化判别函数系数，可以了解每个变量对判别值的相对作用。在判别函数中，系数越大的变量，在该判别函数中的相对作用越大。例如，总资产贡献率、资产负债率和工业成本费用利润率在第一个判别函数中的作用比较大，而流动资产周转率在第二个判别函数中的作用比较大。

表 15-8(6)　标准化判别函数系数

	Function(函数)	
	1	2
总资产贡献率	0.482	0.018
资产负债率	0.132	0.006
流动资产周转率	–0.357	0.972
工业成本费用利润率	0.853	0.172

(8)表 15-8(7)是 Fisher 判别系数。通过 Fisher 判别系数，可建立 Fisher 判别函数，进行手工判别。

表 15-8(7)　分类函数系数

	类型		
	1.00	2.00	3.00
总资产贡献率	5.657	3.762	3.450
资产负债率	8.708	8.419	8.370
流动资产周转率	49.040	73.961	56.518
工业成本费用利润率	16.006	11.003	9.666
(Constant)	–434.219	–388.899	–343.955

Fisher's linear discriminant functions(Fisher 线性判别系数)。

F_1= –434.219+5.657×总资产贡献率+8.708×资产负债率+49.04×流动资产周转率+16.006×工业成本费用利润率

F_2= –388.899+3.762×总资产贡献率+8.419×资产负债率+73.961×流动资产周转率+11.003×工业成本费用利润率

F_3= –343.955+3.450×总资产贡献率+8.370×资产负债率+56.518×流动资产周转率+9.666×工业成本费用利润率

将每个新来的样品数据带入函数中，分别计算出 F_1、F_2 和 F_3，哪个 F 值大就判为哪一类。

(9)表 15-8(8)是利用判别函数看样品数据回判结果。

表 15-8(8)　分类结果(b,c)

		类型	Predicted Group Membership (判别样品所属组别)			Total
			1.00	2.00	3.00	
Original	Count	1.00	20	0	0	20
		2.00	0	20	0	20
		3.00	0	0	20	20

续表

		类型	Predicted Group Membership（判别样品所属组别）			Total
			1.00	2.00	3.00	
Original		Ungrouped cases（未分组）	1	1	0	2
	%	1.00	100.0	0.0	0.0	100.0
		2.00	0.0	100.0	0.0	100.0
		3.00	0.0	0.0	100.0	100.0
		Ungrouped cases（未分组）	50.0	50.0	0.0	100.0
Cross-validated	Count	1.00	20	0	0	20
		2.00	0	20	0	20
		3.00	0	0	20	20
	%	1.00	100.0	0.0	0.0	100.0
		2.00	0.0	100.0	0.0	100.0
		3.00	0.0	0.0	100.0	100.0

表 15-8(8)分成两部分：上面一半(Original)是利用从全部数据得到的判别函数来判断每一个样品的结果(前面三行为判断结果的数目，而后三行为相应的百分比)。

下面一半(Cross-validated)是对每一个样品，都用缺少该样品数据的其他样品数据得到的判别函数来进行判别的结果。

从回判结果看，正确率为 100%，判别效果还是比较好的。另外两个待判的样品(Case)一个被判为第一类，另一个判为第二类。

思考与练习

1．什么是距离和相似系数？

2．系统聚类分析的基本思想是什么？系统聚类分析方法有哪些？

3．判别哪种聚类方法好的标准是什么？

4．什么是判别分析？它与聚类分析是什么关系？

5．Fisher 判别分析方法的思想是什么？

6．判别分析法的前提条件是什么？

7．对某年我国各地区收入结构差异进行聚类和判别分析。

8．对上一年我国各地区房地产经济情况进行判别分析。

9．利用判别法分析各地区农民家庭消费结构。

10．根据中国县(市)社会经济统计年鉴，搜集各大农业主产区相关指标并进行聚类分析。要求变量为：人口、第一产业增加值、第二产业增加值、居民存款余额、工业总产值、中学在校学生数和医院床位数。

11．搜集亚洲各国家三大产业的统计数据资料(数据来源《国际统计年鉴》)，并对其国民经济的贡献率进行聚类分析。要求变量为：第一产业占 GDP 的比重、第二产业占 GDP 的比重和第三产业占 GDP 的比重。

12．搜集我国各省市不同类型的房屋销售的统计数据资料，并进行聚类分析。要求聚类

变量为：别墅公寓、经济适用房、办公楼、商业用房和其他。

13．搜集我国各省市城镇居民平均每人全年家庭收入来源的统计数据资料，并进行聚类分析。要求变量为：国有单位职工收入、集体单位职工收入、其他经济类型单位职工收入、财产性收入和转移性收入。

14．搜集我国西部各省市经济发展水平统计数据，并进行聚类分析。要求：变量为国内生产总值、工业增加值、固定资产投资、居民消费价格指数(%)、外贸进出口额、社会消费品零售总额和一般预算收入，将各省市分为两类。

15．搜集我国各地区教育经费的统计数据资料，通过聚类分析划分为不同类型，并进行判别分析。变量为：国家财政性教育经费、社会团体和公民个人办学经费、社会捐资和集资办学经费、学费和杂费和其他教育经费。将各地区分为三类，其中上海和新疆是待判样本。

16．搜集各地区企业科学研究和开发情况的数据资料并进行判别分析。变量为 R&D 经费内部支出(万元)、基础研究(万元)、应用研究(万元)和试验发展(万元)。要求将各地区分为四类，其中北京和广东是待判样本。

第十六章　主成分分析与因子分析

统计分析实践中，我们经常会搜集到关于研究对象的众多变量和数据。例如，对全国或各个地区的社会经济发展进行描述时会有许多的变量与数据；反映学校教学、科研等整体状况的也有众多的指标，等等。类似这些问题，其共同特点是变量很多，但是这些变量之间有很多又存在着密切的相关关系。对此类问题分析时，我们希望能够找出众多指标的少数“代表”来对所研究的对象进行描述、分析，甚至是评价。这种统计方法实质上是对指标进行降维，一般我们常用的指标降维的方法有主成分分析(Principal Component Analysis)和因子分析(Factor Analysis)等方法，其中主成分分析可以说是因子分析的一个特例。本章的目的就是让我们学会如何建立和使用降维模型，即将多个变量的数据减少到只需要用很少的几个变量来表示。

第一节　主成分分析

在介绍主成分分析之前，先看下面的例子。

我们知道各地区的社会与经济发展水平存在着一定的差异，而可以反映社会经济发展水平的指标有很多，如地区生产总值(亿元)、在岗职工平均工资(元/人)、城镇居民人均可支配收入(元)、地方财政收入(亿元)、全社会固定资产投资(亿元)、社会消费品零售额(亿元)和从业人数(人)等很多指标。

现在我们的问题是如何用和用什么指标反映各地区的社会经济发展水平呢？以及如何反映各地区之间的差异呢？如果直接利用这些众多的指标去分析，会遇到很多的问题，首先众多指标中的每一个指标只能反映社会经济发展的某一个方面，因而用一个指标很难以偏概全。那么能否将这些指标直接综合起来呢？答案是不能，因为这些指标之间存在着相关性，有些指标之间的相关性还很强。表 16-1 就显示出地区生产总值、在岗职工平均工资、城镇居民人均可支配收入、地方财政收入、全社会固定资产投资、社会消费品零售额和从业人数这七个指标之间都存在着显著的相关关系。如果直接用这些指标进行综合分析，则会对那些相关性很强的信息重复计算，人为地夸大它们的作用。

表 16-1　相关系数矩阵(Correlation Matrix)

		在岗职工平均工资	地区生产总值	城镇居民人均可支配收入	地方财政收入	全社会固定资产投资	社会消费品零售额	从业人数
相关系数	在岗职工平均工资(元/人)	1.000	0.711	0.753	0.805	0.621	0.775	0.772
	地区生产总值(万元)	0.711	1.000	0.615	0.934	0.944	0.964	0.968
	城镇居民人均可支配收入(元)	0.753	0.615	1.000	0.691	0.562	0.640	0.620
	地方财政收入(万元)	0.805	0.934	0.691	1.000	0.939	0.917	0.950
	全社会固定资产投资(万元)	0.621	0.944	0.562	0.939	1.000	0.883	0.910
	社会消费品零售额(万元)	0.775	0.964	0.640	0.917	0.883	1.000	0.979
	从业人数(人)	0.772	0.968	0.620	0.950	0.910	0.979	1.000

续表

		在岗职工平均工资	地区生产总值	城镇居民人均可支配收入	地方财政收入	全社会固定资产投资	社会消费品零售额	从业人数
Sig	在岗职工平均工资(元/人)		0.000	0.000	0.000	0.003	0.000	0.000
	地区生产总值(万元)	0.000		0.003	0.000	0.000	0.000	0.000
	城镇居民人均可支配收入(元)	0.000	0.003		0.001	0.008	0.002	0.003
	地方财政收入(万元)	0.000	0.000	0.001		0.000	0.000	0.000
	全社会固定资产投资(万元)	0.003	0.000	0.008	0.000		0.000	0.000
	社会消费品零售额(万元)	0.000	0.000	0.002	0.000	0.000		0.000
	从业人数(人)	0.000	0.000	0.003	0.000	0.000	0.000	

这时就可以用主成分分析和因子分析的方法，将相关性较高的指标综合成一个指标，从而用较少的综合指标来反映原来众多变量中较多的信息，达到降维、简化分析过程的目的。

现在的问题是如何将这些变量用少数几个综合变量来表示呢？这些少数的几个综合变量又包含有多少原来的信息呢？能不能利用找到的综合变量来对不同地区的社会经济发展水平排序呢？这一类数据所涉及的问题可以推广到对企业、对学校等进行分析、排序、判别和分类等问题。

一、主成分分析的降维思路

为了说明如何将指标进行降维，我们可以借助于图示的方式来展现。首先来看一种最简单的情况，即如何从二维变量降到一维。

假设现在只有两个变量，分别由横坐标和纵坐标所代表(如图 16-1 所示)，每个观测值都有相应于这两个坐标轴的两个坐标值。如果这些数据形成一个椭圆形状的点阵(这在变量的二维正态的假定下是可能的)，我们便可得到椭圆的长轴和短轴，我们可以看到在椭圆的短轴方向上数据变化很小，而在长轴的方向上数据变化则很大。当短轴退化到极端的情况成为一点时，所有的数据就都落在了长轴上，这时只用长轴就能够完全解释这些数据的变化，我们只保留一个维度上的信息就可以了。这样，指标由二维到一维的降维工作就自然完成了。

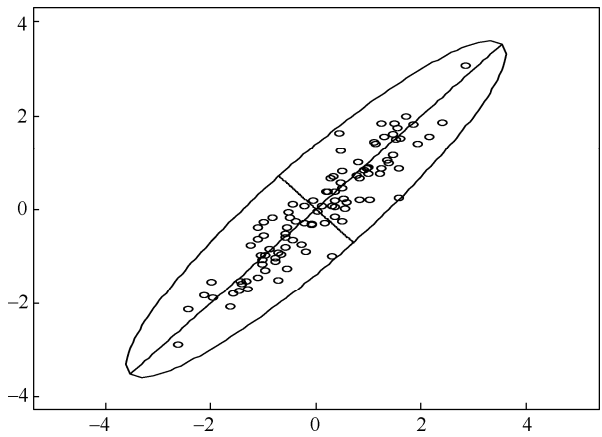


图 16-1　主成分分析的降维

当我们将坐标轴旋转，旋转到坐标轴和椭圆的长短轴平行，那么长轴代表的变量就描述了数据的主要变化，而短轴代表的变量就描述了数据的次要变化。这样我们就可以用长轴变

量反映众多指标中所包含的数据的大部分信息，而忽略短轴所代表的信息(舍去次要的一维)，这样降维就完成了。椭圆(球)的长短轴相差得越大，降维效果越好。

多维变量的情况其实和二维类似，可以用一个高维的椭球，只不过无法直观地看见罢了。我们可以首先把高维椭球的各个主轴找出来，再用能代表较多数据信息的若干个相对较长的轴作为新的综合变量保留，其余的忽略。这样，主成分分析就基本完成了。

注意：和二维情况类似，高维椭球的各主轴之间也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，称为主成分(Principal Component)。正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。

选择越少的主成分同时保留的信息量越多，降维的效果就越好。降维的标准是被选的主成分所代表的主轴的长度之和占了所有主轴长度总和的大部分。有些文献建议，所选的主轴总长度占所有主轴长度之和的大约 85%即可，具体选几个，要看实际情况而定。

对于我们的数据，SPSS 输出的结果如表 16-2 所示。

表 16-2 Total Variance Explained(方差解释表)

主成分	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.886	84.082	84.082	5.886	84.082	84.082
2	0.674	9.629	93.710	0.674	9.629	93.710
3	0.256	3.659	97.369			
4	0.122	1.737	99.106			
5	0.035	0.493	99.599			
6	0.018	0.253	99.852			
7	0.010	0.148	100.000			

Extraction Method：Principal Component Analysis.

这里的 Initial Eigenvalues 就是指七个主轴长度，又称特征值(是变量相关系数矩阵的特征值)。每个特征值占特征值总和的百分比就是该特征值所对应的主成分的方差贡献率。表 16-2 中前两个成分特征值对应的方差累积占了总方差的 93.710%，则累计方差贡献率为 93.710%。后面的特征值的贡献越来越少，一般我们取累计方差贡献率达到 85%左右的前 k 个主成分就可以了，因为它们已经包含了绝大部分的信息。

二、主成分分析的一般模型

主成分或这些主轴用数学公式表达出来就是[①]：

$$F_1 = \mu_{11}x_1 + \mu_{12}x_2 + \cdots + \mu_{1p}x_p$$

$$F_2 = \mu_{21}x_1 + \mu_{22}x_2 + \cdots + \mu_{2p}x_p$$

$$\cdots$$

$$F_p = \mu_{p1}x_1 + \mu_{p2}x_2 + \cdots + \mu_{pp}x_p$$

式中 F_1 就是第一个主成分，F_2 是第二个主成分……

① 这是介绍主成分的一般数学模型，是为了能更好地理解因子分析，所以只要求了解即可。

μ_{ij} 为变量 X 的系数，满足：$\mu_{k1}^2+\mu_{k2}^2+\cdots\mu_{kp}^2=1$。

μ_{ij} 由以下原则来确定：

F_i 与 F_j 互斥；

F_1 是 $x_1\cdots x_p$ 的线性组合，是所有主成分中方差最大的；

F_2 也是 $x_1\cdots x_p$ 的线性组合，是所有主成分中方差第二大的，仅小于 F_1 的方差；

F_3 也是 $x_1\cdots x_p$ 的线性组合，是所有主成分中方差第三大的，仅小于 F_1、F_2 的方差；

以此类推……

这时称：F_1 是第一主成分，F_2 是第二主成分……

这些方差的大小就对应于每一个主轴的长度。

第二节　因 子 分 析

一、因子分析的目的

因子分析是主成分分析的推广和发展。既然有了主成分分析，为什么还要进行因子分析呢？由主成分分析的模型可知：

$$\begin{aligned}F_1&=\mu_{11}x_1+\mu_{12}x_2+\cdots+\mu_{1p}x_p\\F_2&=\mu_{21}x_1+\mu_{22}x_2+\cdots+\mu_{2p}x_p\\&\vdots\\F_p&=\mu_{p1}x_1+\mu_{p2}x_2+\cdots+\mu_{pp}x_p\end{aligned}$$

首先，在主成分分析中，每个主成分与成分系数 μ_{ij} 之间的关系不明确；此外主成分分析不能表达每个变量 x_i 与提出来的主成分 F_i 的关系，因此就要进行因子分析。因子分析模型为：

$$\begin{aligned}x_1&=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\varepsilon_1\\x_2&=a_{21}F_1+a_{22}F_2+\cdots+a_{2p}F_P+\varepsilon_2\\&\vdots\\x_p&=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\varepsilon_p\end{aligned}$$

式中，F_i 称为公共因子；a_{ij} 称为因子载荷。

因子载荷的统计意义就是第 i 个变量与第 j 个公共因子的相关系数，即表示变量 x_i 依赖于 F_j 的分量(比重)。

如上例中，我们利用主成分的方法提取出两个公因子：F_1 和 F_2(如表 16-3 所示)。

表 16-3　因子载荷表

	Component(因子)	
	1	2
地方财政收入(万元)	0.977	−0.055
从业人数(人)	0.974	−0.152
社会消费品零售额(万元)	0.967	−0.116

续表

	Component（因子）	
	1	2
地区生产总值（万元）	0.965	–0.206
全社会固定资产投资（万元）	0.923	–0.286
在岗职工平均工资（元/人）	0.841	0.400
城镇居民人均可支配收入（元）	0.746	0.591

Extraction Method：Principal Component Analysis（提取方法：主成分分析）。

根据表 16-3，可以写出因子模型：

$$\text{地方财政收入}=0.977F_1-0.055F_2$$

其含义为地方财政收入与公因子 F_1 的相关程度高达 0.977，而与公因子 F_2 是负相关且相关程度只有 0.055。

同理，可以写出其他关系式：

$$\text{从业人数}=0.974F_1-0.152F_2$$

$$\vdots$$

从因子载荷表中我们发现，所有变量都与公因子 F_1 有很高的载荷系数，而与公因子 F_2 的载荷系数相对都比较低。那么该如何解释公因子 F_1 和 F_2 呢？为了更好地解释公因子的含义，我们通常需要进行因子旋转。所谓对公因子更好地进行解释，就是使每个变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小。

二、因子旋转

因子旋转就是使每个变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小，即尽量使一个变量的信息集中在某一个公因子上，目的是进行因子命名。这种变换因子载荷的方法被称为因子轴的旋转。因子旋转的方法有很多种，常用的为方差最大正交旋转。对所提取公因子进行旋转之后的结果如表 16-4 所示。

表 16-4 Rotated Component Matrix①（旋转因子矩阵）

	Component（因子）	
	1	2
全社会固定资产投资（万元）	0.926	0.277
地区生产总值（万元）	0.916	0.366
从业人数（人）	0.894	0.417
社会消费品零售额（万元）	0.868	0.442
地方财政收入（万元）	0.842	0.499
城镇居民人均可支配收入（元）	0.290	0.907
在岗职工平均工资（元/人）	0.475	0.801

① Extraction Method：Principal Component Analysis. Rotation Method：Varimax with Kaiser Normalization（提取方法：主成分分析方法，旋转方法：方差最大正文旋转）。

这里，第一个因子主要和全社会固定资产投资、地区生产总值、从业人数、社会消费品零售额、地方财政收入有很强的正相关；而第二个因子主要和城镇居民人均可支配收入、在

岗职工平均工资有很强的正相关。因此可以给第一个因子命名为“总量因子”，而给第二个因子命名为“人均因子”。从这里可以看出，因子分析的结果比主成分分析解释性更强。

三、因子得分

对每个因子命名之后，因子分析并没有结束。因为我们不仅想知道第一个因子是总量因子，第二个因子是人均因子，还要知道我们考察的各地区在总量和人均方面的情况到底如何。这就需要计算因子得分，即各地区在总量因子和人均因子上的得分究竟是多少。

因子得分的计算也是根据原始各变量的数据计算而成，是各变量的线性组合，可以表示为：

$$\begin{aligned} f_1 &= \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p \\ f_2 &= \beta_{21}x_1 + \beta_{22}x_2 + \cdots + \beta_{2p}x_p \\ &\vdots \\ f_m &= \beta_{m1}x_1 + \beta_{m2}x_2 + \cdots + \beta_{mp}x_p \end{aligned}$$

上式称为因子得分函数，f_1 就是因子 F_1 的得分，f_2 就是因子 F_2 的得分。用它可计算每个样品(在本例中就是每个地区)的公因子得分。估计因子得分的方法很多，常用的是线性回归的方法，利用线性回归法得到的因子得分系数如表 16-5 所示。

表 16-5　Component Score Coefficient Matrix(因子得分系数矩阵)

	Component(因子)	
	1	2
在岗职工平均工资(元/人)	–0.212	0.573
地区生产总值(万元)	0.307	–0.163
城镇居民人均可支配收入(元)	–0.384	0.799
地方财政收入(万元)	0.183	0.025
全社会固定资产投资(万元)	0.367	–0.265
社会消费品零售额(万元)	0.233	–0.052
从业人数(人)	0.263	–0.095

可以根据因子得分系数，计算出每个地区的第一个因子和第二个因子的得分大小，即计算出每个地区的因子得分 f_1 和 f_2：

$$f_1 = 0.307x_1 + 0.183x_2 + 0.367x_3 + 0.233x_4 + 0.263x_5 - 0.384x_6 - 0.212x_7$$

$$f_1 = -0.163x_1 + 0.025x_2 - 0.265x_3 - 0.052x_4 - 0.095x_5 + 0.799x_6 + 0.573x_7$$

我们可以根据这两个函数分别计算出每个地区的两个因子得分，对各地区分别按照总量因子和人均因子排序。也可以以旋转后的每个因子的方差贡献率为权数，进行加权综合，计算出每个地区的总得分，以此进行比较排序。

$$\begin{aligned} \text{总得分} &= f_1 \times f_1\text{的方差贡献率} + f_2 \times f_2\text{的方差贡献率} \\ &= f_1 \times 5.886 + f_2 \times 0.674 \end{aligned}$$

四、主成分和因子分析的一些注意事项

可以看出，因子分析和主成分分析都依赖于原始变量，并在降维后利用公因子反映原始变量的信息，所以原始变量的选择很重要。另外，如果原始变量都本质上独立，那么降维就

可能失败，这是因为很难把很多独立变量用少数综合的变量概括。变量之间的相关程度越高，降维的效果就越好。

在面对实际问题进行因子分析时，并不一定都会得到如我们例子那样清楚的结果。这与问题的性质，选取的原始变量以及数据的特征等都有关系。

在用因子得分进行排序时要特别小心，特别是对于敏感问题。由于原始变量不同，因子的选取不同，排序可以很不一样。

五、因子分析的 SPSS 实现与输出结果解读

我们仍以本章例子来说明利用 SPSS 软件进行因子分析的过程。

SPSS 中因子分析的操作方法如下。

第一步：单击【Analyze】→【Data Reduction】→【Factor Analysis】，进入因子分析。

第二步：将地区生产总值等七个变量选入【Variables】。

第三步：单击【Descriptives】进入，选择【Correlations Matrix】中的“Coefficients(输出相关系数矩阵)”和“KMO and Bartletts test of Sphericity(KMO 检验和巴特利球形检验)”。

第四步：单击【Extraction】进入，在【Display】中选择“Scree plot”（输出碎石图)。

第五步：单击【Rotation】进入，选择【Method】中的 Varimax(方差最大化)。

第六步：单击【Scores】进入，选择【Save as Variables】(因子得分就会作为变量存在数据中的附加列上)和【Display factor score cofficient matrix】(输出因子得分表)。

第七步：如果想对输出结果排序，可单击【Option】进入，选择【Sort by size】(排序)即可。

第八步：如果要自己确定提取的公因子个数，可在【Extraction】中的【Extract】中选择“Number of Factor”，然后输入数字。

第九步：单击【OK】按钮即可。

(一)KMO 测度和巴特利特球体检验

因子分析的目的是简化数据，因此使用因子分析的前提条件是研究变量之间应该有较强的相关关系。如果变量之间的相关程度很低，则不适合进行因子分析。SPSS 提供了几种帮助判断观测数据是否适合做因子分析的方法，即 KMO(Kaiser-Meyer-Olkin Measure of Sampling Adequacy)测度和巴特利特球体检验(Bartletts Test of Sphericity)，其检验计算的结果如表 16-6 所示。

表 16-6 KMO and Bartlett's Test(KMO 和巴特利特检验)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.820
Bartlett's Test of Sphericity	Approx. Chi-Square	194.515
	df	21
	Sig.	0.000

KMO 测度值的变化范围从 0 到 1。KMO 值较小时，表明观测变量不适合做因子分析。通常判断标准为：0.9 以上，非常好；0.8 以上，好；0.7，一般；0.6，差；0.5，很差；0.5 以下，不适合做因子分析[①]。

① 郭志刚. 社会统计分析方法-SPSS 软件应用. 北京：中国人民大学出版社，1999.

巴特利特球体检验。该检验是从检验整个相关矩阵出发，其零假设为相关矩阵是单位阵(即该矩阵没有显著的相关关系)。如果不能拒绝零假设，则不适宜进行因子分析。

从 SPSS 输出结果可知，KMO 测度值已达到 0.8 以上，并且巴特利特球体检验的 P 值=0.000，拒绝零假设，因此说明该观测数据适合做因子分析。

(二)公因子方差表

提取出来的公因子对每个变量的解释程度到底有多大呢？可从公因子方差表 16-7 中得知。

表 16-7　Communalities(公因子方差表)

	Initial	Extraction
在岗职工平均工资(元/人)	1.000	0.707
地区生产总值(万元)	1.000	0.931
城镇居民人均可支配收入(元)	1.000	0.556
地方财政收入(万元)	1.000	0.954
全社会固定资产投资(万元)	1.000	0.852
社会消费品零售额(万元)	1.000	0.935
从业人数(人)	1.000	0.949

Extraction Method：Principal Component Analysis.

公因子方差表告诉了我们提取出来的公因子对每个变量的解释程度，如公因子对变量在岗职工平均工资的解释程度为 70.7%，对变量地区生产总值的解释程度为 93.1%，等等。提取出来的公因子对变量的解释程度越高，说明提取的公因子包含原有变量的信息量越大。

(三)公因子的提取

提取公因子的方法有很多，如主成分方法、主轴因子法、普通二乘法等，其中最常用的是利用主成分的方法来提取公因子。提取公因子个数的标准通常有三个：一是根据特征根≥1 来选取(这是 SPSS 默认的标准)；第二种是根据分析人员计划要提取的信息量，即根据累计方差贡献率确定；此外还有一种是使用者直接规定公因子的个数来选取。当 SPSS 选取的公因子个数不能满足研究的需要时，则可以由用户直接规定公因子的个数来选取(如表 16-8 所示)。

表 16-8　Total Variance Explained(总方差解释表)

主成分	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.886	84.082	84.082	5.886	84.082	84.082
2	.674	9.629	93.710			
3	.256	3.659	97.369			
4	.122	1.737	99.106			
5	.035	0.493	99.599			
6	.018	0.253	99.852			
7	.010	0.148	100.000			

Extraction Method：Principal Component Analysis.

表 16-8 的输出结果说明，只提取了一个公因子就可解释原有信息的 84.08%。按照公因子提取的一般标准，这一个公因子包含的信息量已基本上满足要求。但是在实际分析中，仅

提取一个公因子对分析结果不好解释，因此我们需要提取两个公因子。这时可以重新运行因子分析，在 Extraction 中选择“Number of Factor”，然后输入 2 即可。重新运行因子分析输出结果如表 16-9 所示。

表 16-9 Total Variance Explained(总方差解释表)

主成分	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.886	84.082	84.082	5.886	84.082	84.082
2	.674	9.629	93.710	.674	9.629	93.710
3	.256	3.659	97.369			
4	.122	1.737	99.106			
5	.035	0.493	99.599			
6	.018	0.253	99.852			
7	.010	0.148	100.000			

Extraction Method：Principal Component Analysis.

在提取两个公因子后，累计方差贡献率已高达 93.71%，即这两个公因子已解释原有信息的 93%以上。当然两个公因子对每个变量的解释程度也都有提高(如表 16-10 所示，公因子方差表)。

表 16-10 Communalities(公因子方差表)

	Initial	Extraction
在岗职工平均工资(元/人)	1.000	0.867
地区生产总值(万元)	1.000	0.974
城镇居民人均可支配收入(元)	1.000	0.906
地方财政收入(万元)	1.000	0.957
全社会固定资产投资(万元)	1.000	0.934
社会消费品零售额(万元)	1.000	0.949
从业人数(人)	1.000	0.973

Extraction Method：Principal Component Analysis.

特征值的贡献还可以从 SPSS 的所谓碎石图(如图 16-2 所示)中看出。

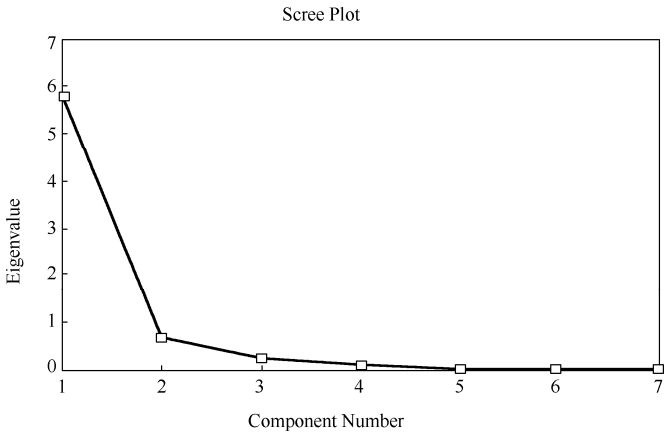


图 16-2 碎石图

（四）因子载荷矩阵

提取出来的两个公因子与原变量之间的关系可从因子载荷矩阵中得到（如表 16-11 所示）：

表 16-11　Component Matrix（因子载荷矩阵）

	Component（因子）	
	1	2
在岗职工平均工资（元/人）	0.841	0.400
地区生产总值（万元）	0.965	–0.206
城镇居民人均可支配收入（元）	0.746	0.591
地方财政收入（万元）	0.977	–0.055
全社会固定资产投资（万元）	0.923	–0.286
社会消费品零售额（万元）	0.967	–0.116
从业人数（人）	0.974	–0.152

Extraction Method：Principal Component Analysis.

从因子载荷矩阵中可知，第一个公因子与各变量的载荷都很高，而第二公因子与各变量的载荷都较低，所以我们很难直接去决定公因子一代表哪些变量，公因子二表示哪些变量。因此需要进行公因子旋转。

（五）公因子旋转

前面我们介绍了公因子旋转的目的是使每个变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小。选择常用的方差最大旋转方法，计算结果如表 16-12 所示。

表 16-12　Rotated Component Matrix（旋转后因子载荷矩阵）

	Component（因子）	
	1	2
全社会固定资产投资（万元）	0.926	0.277
地区生产总值（万元）	0.916	0.366
从业人数	0.894	0.417
社会消费品零售额（万元）	0.868	0.442
地方财政收入（万元）	0.842	0.499
城镇居民人均可支配收入（元）	0.290	0.907
在岗职工平均工资（元/人）	0.475	0.801

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

因子旋转后，明显可见第一个公因子与全社会固定资产投资、地区生产总值、从业人数、社会消费品零售总额和地方财政收入的因子载荷系数很高，而这些指标都是从总量上来反映各地区的经济发展水平的，因此可称第一公因子为经济发展总量因子；第二公因子与城镇居民人均可支配收入和在岗职工平均工资的载荷系数很高，这些指标是从平均水平来说明地区经济发展水平的，因此可称第二公因子为经济发展人均水平因子。在实际分析应用中，对公因子的命名一定要结合研究的具体问题来确定。

(六)因子得分

SPSS 中会输出因子得分系数矩阵，同时也会将每个观测值相应的因子得分计算出来，即将每个地区的公因子一和公因子二的因子得分分别计算出来，并作为变量存入到数据文件中。

第三节　应用案例：我国工业企业经济效益评价

工业企业经济效益评价主要是对工业生产经营活动的各个方面和全部过程的经济效益状况进行全面、客观、定量的反映和科学考察。对全国及地区工业企业的经济效益进行评价有着重要的意义。这种综合评价有利于正确、全面地反映工业企业经济效益的实际水平，分析其变化速度和变动趋势，发现问题找出差距。对工业企业经济效益评价的方法有很多，在此主要应用因子分析的方法对企业的经济效益进行评价。

对工业企业经济效益评价首先要确定评价的指标体系，在实际中有很多关于工业企业经济效益评价的考核指标，它们从不同方面和不同角度来反映工业企业经济效益，本例中选用了由总资产贡献率、资本保值增值率、资产负债率、流动资产周转率、成本费用利用率、全员劳动生产率、产品销售率七项指标组成的指标体系(如表 16-13 所示)。

表 16-13　****年国有及规模以上非国有工业企业经济效益数据

	总资产贡献率%	资本保值增值率%	资产负债率%	流动资产周转率%	成本费用利用率%	全员劳动生产率元/人	产品销售率%
北　京	9.85	33.64	51.61	1.80	6.41	12603.82	97.86
天　津	13.36	43.18	56.70	2.31	8.19	11830.13	98.88
河　北	13.50	71.79	62.32	2.51	6.80	8989.95	98.52
山　西	11.24	105.48	66.72	1.64	7.24	6682.08	97.75
内蒙古	10.77	106.92	61.19	2.12	6.46	9880.44	98.55
辽　宁	9.08	55.25	58.92	1.93	5.00	9106.21	98.09
吉　林	10.78	73.65	60.36	2.00	6.04	9906.35	97.62
黑龙江	25.02	45.96	57.74	1.96	27.58	12216.94	98.11
上　海	12.60	16.69	50.25	2.02	7.80	14343.11	99.02
江　苏	11.51	63.48	62.20	2.43	4.77	10347.46	97.86
浙　江	13.25	42.44	57.83	2.20	6.00	7412.79	97.63
安　徽	10.75	79.87	60.69	1.96	5.22	7026.82	98.51
福　建	11.82	39.74	52.78	2.30	6.07	7509.05	97.59
江　西	10.33	102.16	65.20	2.08	3.61	6194.73	98.01
山　东	15.46	71.24	58.50	2.79	7.11	10031.18	97.93
河　南	12.30	81.11	64.51	2.23	5.70	7135.39	98.52
湖　北	8.87	76.66	58.04	1.80	5.96	8451.54	98.45
湖　南	12.94	95.62	63.57	2.15	4.96	7551.37	99.83
广　东	11.06	25.78	58.11	2.35	5.12	8703.41	97.27
广　西	12.00	106.81	64.74	1.97	7.20	6958.76	97.82
海　南	13.24	28.87	54.49	1.69	8.52	10675.68	96.46
重　庆	10.78	67.53	60.53	1.76	5.77	6523.75	97.83
四　川	8.83	69.16	63.28	1.62	4.66	7436.88	98.8
贵　州	10.70	70.17	66.07	1.47	5.90	6562.87	97.08
云　南	19.40	39.03	54.94	1.49	12.05	13877.17	99.36
西　藏	13.10	–14.53	25.72	0.82	18.19	7868.85	94.02

续表

	总资产贡献率%	资本保值增值率%	资产负债率%	流动资产周转率%	成本费用利用率%	全员劳动生产率元/人	产品销售率%
陕　西	12.38	92.29	63.44	1.51	11.31	7545.72	97.59
甘　肃	9.18	91.87	60.93	1.69	4.51	6553.78	97.88
青　海	8.91	134.41	71.03	1.16	12.83	9615.11	97.14
宁　夏	7.57	86.92	64.37	1.57	3.24	6145.49	96.71
新　疆	16.73	69.52	57.13	2.15	20.95	16044.21	98.25

在上面七个指标中，只有资产负债率为逆指标，因此在进行分析之前要将其转变为正指标，即先将其求倒数，然后再进行因子分析。

首先来考察这七个工业经济效益评价考核指标的相关性，由前面知道，因子分析的前提条件是研究的变量之间应该具有相关性。首先进行相关分析(如表 16-14 所示)和 KMO 测度与巴特利特球体检验(如表 16-15 所示)。

表 16-14　Correlation Matrix(相关系数矩阵)

		总资产贡献率	资本保值增值率	资产负债率的倒数	流动资产周转率	成本费用利用率	全员劳动生产率	产品销售率
相关系数	总资产贡献率	1.000	–0.324	0.147	0.182	0.766	0.549	0.161
	资本保值增值率	–0.324	1.000	–0.688	0.023	–0.252	–0.392	0.310
	资产负债率的倒数	0.147	–0.688	1.000	–0.401	0.379	0.117	–0.629
	流动资产周转率	0.182	0.023	–0.401	1.000	–0.264	0.163	0.514
	成本费用利用率	0.766	–0.252	0.379	–0.264	1.000	0.531	–0.213
	全员劳动生产率	0.549	–0.392	0.117	0.163	0.531	1.000	0.263
	产品销售率	0.161	0.310	–0.629	0.514	–0.213	0.263	1.000
Sig.	总资产贡献率		0.038	0.215	0.163	0.000	0.001	0.194
	资本保值增值率	0.038		0.000	0.452	0.086	0.015	0.045
	资产负债率的倒数	0.215	0.000		0.013	0.018	0.266	0.000
	流动资产周转率	0.163	0.452	0.013		0.076	0.190	0.002
	成本费用利用率	0.000	0.086	0.018	0.076		0.001	0.125
	全员劳动生产率	0.001	0.015	0.266	0.190	0.001		0.076
	产品销售率	0.194	0.045	0.000	0.002	0.125	0.076	

表 16-15　KMO and Bartlett's Test(KMO 测度和巴特利特球体检验)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.459
Bartlett's Test of Sphericity	Approx. Chi-Square	120.247
	df	21
	Sig.	0.000

由相关系数矩阵可知，在一些评价指标之间存在有显著的相关性，如总资产贡献率与资本保值增值率、工业成本费用利润率和全员劳动生产率都存在显著的相关关系；资产负债率与资本保值增值率、流动资产周转率和工业成本费用利润率之间也存在着显著的相关关系。显然，直接加总来评价必然会产生评价信息的重复。

再看 KMO 测度和巴特利特球体检验的结果，KMO 测度值只有 0.459，不适合进行因子分析，但是巴特利特球体检验却通过了相关性的显著性检验。那么是否能够进行因子分析呢？统计测度值仅是提供了人们判断参考的依据，并不是绝对的。因此是否适合进行因子分

析，关键是其结果是否符合实际情况，是否能够进行比较合理的解释。鉴于此我们还是可以继续进行因子分析。

其次，利用主成分的方法提取公因子，得到总方差解释表(如表 16-16 所示)。

表 16-16 Total Variance Explained(总方差解释表)

主成分	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.784	39.768	39.768	2.784	39.768	39.768
2	2.120	30.291	70.059	2.120	30.291	70.059
3	0.999	14.272	84.331			
4	0.535	7.637	91.968			
5	0.309	4.407	96.376			
6	0.194	2.765	99.141			
7	0.060	0.859	100.000			

Extraction Method：Principal Component Analysis.

SPSS 按照特征值大于 1 的标准提取了两个公因子，两个公因子的累积方差贡献率只有 70.06%，比较低，而如果提取 3 个公因子的话，其累积方差贡献率就会达到 84.33%。因此利用给出提取公因子个数的方法，提取 3 个公因子，输出结果如表 16-17 所示。

表 16-17 Total Variance Explained(总方差解释表)

主成分	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.784	39.768	39.768	2.784	39.768	39.768
2	2.120	30.291	70.059	2.120	30.291	70.059
3	.999	14.272	84.331	0.999	14.272	84.331
4	.535	7.637	91.968			
5	.309	4.407	96.376			
6	.194	2.765	99.141			
7	.060	.859	100.000			

Extraction Method：Principal Component Analysis.

提取 3 个公因子后，这 3 个公因子对每个变量的解释程度基本上在 80%以上(如表 16-18 所示)。

表 16-18 Communalities(公因子方差表)

	Initial	Extraction
总资产贡献率	1.000	0.837
资本保值增值率	1.000	0.922
资产负债率的倒数	1.000	0.881
流动资产周转率	1.000	0.800
成本费用利用率	1.000	0.951
全员劳动生产率	1.000	0.714
产品销售率	1.000	0.798

Extraction Method：Principal Component Analysis.

分析碎石图可以看出因子 1 与因子 2、因子 2 与因子 3、因子 3 与因子 4 之间的特征值之差比较大。而因子 4、5、6、7 之间的特征值差值均比较小。也可得出保留三个因子将能涵盖绝大部分信息。

接下来要确定公因子的含义。要先看提取出来的 3 个公因子载荷矩阵(成分矩阵)。但是直接从公因子载荷矩阵(如表 16-19 所示)上看每个因子中各原始变量的系数没有明显的差别，即每个因子的含义不明显，因此需要进行因子旋转，使载荷系数向 0 和 1 两极分化。经过方差最大旋转后得到换转后的因子载荷矩阵(如表 16-20 所示)。

表 16-19　Component Matrix(a)(因子载荷矩阵)

	Component(因子)		
	1	2	3
成本费用利用率	0.805	0.254	0.488
资产负债率的倒数	0.769	–0.472	–0.259
资本保值增值率	–0.730	0.057	0.621
总资产贡献率	0.640	0.625	0.193
产品销售率	–0.445	0.774	–0.025
流动资产周转率	–0.314	0.660	–0.516
全员劳动生产率	0.553	0.636	–0.066

Extraction Method：Principal Component Analysis.

表 16-20　Rotated Component Matrix(旋转后的因子载荷矩阵)

	Component(因子)		
	1	2	3
总资产贡献率	0.901	0.125	–0.100
成本费用利用率	0.895	–0.377	–0.085
全员劳动生产率	0.751	0.312	–0.232
流动资产周转率	0.000	0.894	–0.008
产品销售率	0.164	0.746	0.464
资本保值增值率	–0.235	–0.052	0.929
资产负债率	0.141	–0.476	–0.797

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization.

旋转后的因子意义非常明确：公因子 1 代表了总资产贡献率、成本费用利润率和全员劳动生产率；公因子 2 代表了流动资产周转率和产品销售率；公因子 3 代表了资本保值增值率和资产负债率。在因子分析中对因子的解释或说明是不容易的，根据这些变量的原始含义称公因子 1 为资产运作及获利因子，公因子 2 被暂且称为经营效率因子，公因子 3 被称为发展潜力因子。

最后利用提取出来的这 3 个公因子对各地区的工业经济效益进行评价，即通过计算因子得分(因子得分系数矩阵如表 16-21 所示)，了解各地区在每个因子上的总体水平(如表 16-22 所示)。

表 16-21 Component Score Coefficient Matrix（因子得分系数矩阵）

	Component（因子）		
	1	2	3
总资产贡献率	0.414	0.026	0.073
资本保值增值率	0.069	–0.241	0.627
资产负债率的倒数	–0.044	–0.115	–0.421
流动资产周转率	–0.078	0.565	–0.226
成本费用利用率	0.457	–0.293	0.206
全员劳动生产率	0.298	0.189	–0.096
产品销售率	0.107	0.341	0.178

Extraction Method：Principal Component Analysis.

Rotation Method：Varimax with Kaiser Normalization. Component Scores.

表 16-22 因子得分计算结果与评价结果

地区	公因子 1 得分	公因子 2 得分	公因子 3 得分	未加权综合因子	加权综合因子
北 京	–.099	.350	–1.039	–.788	–8.178
天 津	.438	1.243	–.609	1.072	46.378
河 北	.015	1.098	–.004	1.109	33.804
山 西	–.314	–.797	1.169	.058	–19.945
内蒙古	–.092	.356	.777	1.041	18.192
辽 宁	–.641	.326	–.345	–.660	–20.521
吉 林	–.273	.152	–.014	–.135	–6.452
黑龙江	3.548	–.455	.519	3.613	134.748
上 海	.588	1.226	–1.299	.514	41.958
江 苏	–.317	1.024	–.411	.296	12.557
浙 江	–.374	.490	–.670	–.554	–9.597
安 徽	–.560	.184	.367	–.008	–11.445
福 建	–.577	.582	–.983	–.978	–19.339
江 西	–.858	.071	.729	–.059	–21.594
山 东	.270	1.337	–.361	1.247	46.112
河 南	–.357	.572	.390	.605	8.711
湖 北	–.553	–.005	.243	–.315	–18.664
湖 南	–.114	.857	.894	1.637	34.180
广 东	–.669	.831	–1.264	–1.101	–19.457
广 西	–.252	–.312	.984	.420	–5.419
海 南	.147	–.433	–.988	–1.274	–21.359
重 庆	–.627	–.292	.152	–.767	–31.604
四 川	–.715	–.043	.384	–.374	–24.259
贵 州	–.623	–.930	.376	–1.178	–47.591
云 南	1.924	.269	–.010	2.184	84.538
西 藏	.238	–3.418	–3.374	–6.554	–142.212
陕 西	.242	–1.103	1.024	.163	–9.166
甘 肃	–.852	–.495	.608	–.740	–40.233
青 海	0.317	–1.971	2.050	0.396	–17.843
宁 夏	–1.303	–.962	.393	–1.871	–75.334
新 疆	2.441	0.247	.313	3.001	109.036

有了各个因子得分我们可以计算综合因子得分来进行各地区的工业企业经济效益评价。计算综合因子得分的方法有两种：一种方法是直接将各地区的每个因子的得分相加，即未进行加权的综合因子得分；另一种是以每个公因子的方差贡献率作为权数，进行加权计算的综合因子得分。

在本例中的两种方法计算出的综合因子得分有一定的差异，但不是很大。从计算结果看，排在前三名的是黑龙江、新疆和云南。

思考与练习

1．什么是主成分分析？它的基本思路为何？

2．什么是方差贡献率？方差贡献率在主成分分析中的作用是什么？

3．什么是因子分析？它与主成分分析的区别是什么？

4．因子分析方法的思想是什么？

5．简述因子分析方法的基本步骤。

6．试利用因子分析方法对上一年我国各地区经济效益情况进行综合分析。

7．根据《世界国际竞争力报告》评价不同国家经济发展状况。

8．对我国各地区的文化指标进行综合的分析与评价。

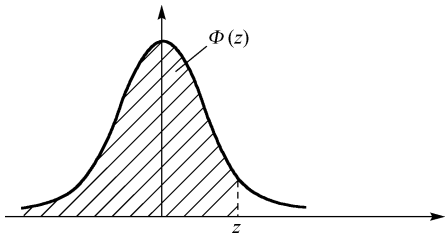
9．搜集数据资料，并对我国各省市、直辖市、自治区工业和生活污染物排放量进行因子分析。变量为：x_1—生活污水排放量(万吨)、x_2—生活二氧化硫排放量(吨)、x_3—生活烟尘排放量(吨)、x_4—工业固体废物排放量(万吨)、x_5—工业废气排放量(亿标立方米)、x_6—工业废水排放量(万吨)。

10．搜集数据资料，并对我国部分地区农村家庭进行人均消费的因子分析。变量为：x_1—食品支出比重、x_2—衣着支出比重、x_3—居住支出比重、x_4—家庭设备及服务支出比重、x_5—医疗保险支出比重、x_6—交通和通讯支出比重、x_7—文教娱乐用品和服务支出比重。

11．搜集数据资料，并对北京市部分地区社会经济发展水平进行因子分析。变量为：x_1—国内生产总值、x_2—地方财政收入、x_3—非农业比重、x_4—人均地方财政收入、x_5—农民人均纯收入、x_6—乡镇企业职工平均工资水平、x_7—每万人拥有医院卫生院技术人员数。

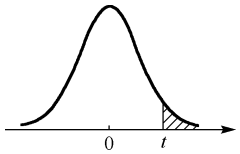
附录 A　常用统计表

表 1　标准正态分布表



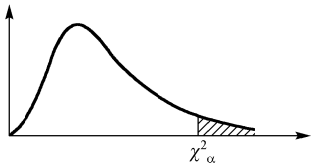
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9278	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9648	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9700	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

表 2　t 分布表



自由度	上　单　侧				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.603	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

表 3　χ^2 分布表

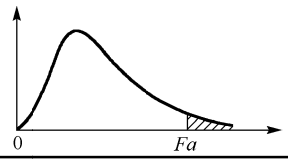


自由度	上 单 侧									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	$392\,704\times10^{-10}$	$157\,088\times10^{-9}$	$982\,069\times10^{-9}$	$393\,214\times10^{-8}$	0.0157908	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720	4.60517	5.99147	7.37776	9.2104	10.9566
3	0.0717212	0.114832	0.215795	0.351846	0.584375	6.25139	7.81473	9.34840	11.3449	12.8381
4	0.206990	0.297110	0.484419	0.710721	1.063623	7.77944	9.48773	11.1433	13.2767	14.8602
5	0.411740	0.554300	0.831211	1.145476	1.61031	9.23635	11.0705	12.8325	15.0863	16.7496
6	0.675727	0.872085	1.237347	1.63539	2.20413	10.6446	12.5916	14.4494	16.8119	18.5476
7	0.989265	1.239043	1.68987	2.16735	2.83311	12.0170	14.0671	16.0128	18.4753	20.2777
8	1.344419	1.646482	2.17973	2.73264	3.48954	13.3616	15.5073	317.5346	20.0902	21.9550
9	1.734926	2.087912	2.70039	3.32511	4.16816	14.6837	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55821	3.24697	3.94030	4.86518	15.9871	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.05347	3.81575	4.57481	5.57779	17.2750	19.6751	21.9200	24.7250	26.7569
12	3.07382	3.57056	4.40379	5.22603	6.30380	18.5494	21.0261	23.3367	26.2170	28.2995
13	3.56503	4.10691	5.00874	5.89186	7.04150	19.8119	22.3621	24.7356	27.6883	29.8194
14	4.07468	4.66043	5.62872	6.57063	7.78953	21.0642	23.6848	26.1190	29.1413	31.3193

续表

自由度	上 单 侧									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
15	4.60094	5.22935	6.26214	7.26094	8.54675	22.3072	24.9958	27.4884	30.5779	32.8013
16	5.14224	5.81221	6.90766	7.96164	9.31223	23.5418	26.2962	28.8454	31.9999	34.2672
17	5.69724	6.40776	7.56418	8.67176	10.0852	24.7690	27.5871	30.1910	33.4087	35.7185
18	6.26481	7.01491	8.23075	9.39046	10.8649	25.9894	28.8693	31.5264	34.8053	37.1564
19	6.84398	7.63273	8.90655	10.1170	11.6509	27.2036	30.1435	32.8523	36.1908	38.5822
20	7.43386	8.26040	9.59083	10.8508	12.4426	28.4120	31.4104	34.1696	37.5662	39.9968
21	8.03366	8.89720	10.28293	11.5913	13.2396	29.6151	32.6705	35.4789	38.9321	41.4010
22	8.64272	9.54249	10.9823	12.3380	14.0415	30.8133	33.9244	36.7807	40.2894	42.7958
23	9.26042	10.19567	11.6885	13.0905	14.8479	32.0069	35.1725	38.0757	41.6384	44.1813
24	9.88623	10.8564	12.4011	13.8484	15.6587	33.1963	36.4151	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	16.4734	34.3816	37.6525	40.6465	44.3141	46.9278
26	11.1603	12.1981	13.8439	15.3791	17.2919	35.5631	38.8852	41.9232	45.6417	48.2899
27	11.8076	12.8786	14.5733	16.1513	18.1138	36.7412	40.1133	43.1944	46.9630	49.6449
28	12.4613	13.5648	15.3079	16.9279	18.9392	37.9159	41.3372	44.4607	48.2782	50.9933
29	13.1211	14.2565	16.0471	17.7083	19.7677	39.0875	42.5569	45.7222	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4926	20.5992	40.2560	43.7729	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4331	26.5093	29.0505	51.8050	55.7585	59.3417	63.6907	66.7659
50	27.9907	29.7067	32.3574	34.7642	37.6886	63.1671	67.5048	71.4202	76.1539	79.4900
60	35.5346	37.4848	40.4817	43.1879	46.4589	74.3970	79.0819	83.2976	88.3794	91.9517
70	43.2752	45.4418	48.7576	51.7393	55.3290	85.5271	90.0012	95.0231	100.425	104.215
80	51.1720	53.5400	57.1532	60.3915	64.2778	96.5782	101.8	106.629	112.329	116.321
90	59.1963	61.7541	65.6466	69.1260	73.2912	107.565	113.145	118.136	124.116	128.299
100	67.3276	70.0648	74.2219	77.9295	82.3581	118.498	124.342	129.561	135.807	140.169

表 4　F 分布表



0.05

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	16.1.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13

续表

0.05

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

续表

0.01

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4 052	4 999.5	5 403	5 625	5 764	5.859	5.928	5 982	6 022	6 056	6 106	6 157	6 209	6 235	6 261	6 287	6 313	6 339	6 366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.06
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21

续表

0.01

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

0.025

第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1 001	1 006	1 010	1 014	1 018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.48	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.98	5.52	5.29	5.21	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33

续表

0.025																			
第二自由度	第一自由度																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	4.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.10	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

表 5 随机数表

63271	59986	71744	51102	15141	80714	58683	93108	13554	79945
88547	09896	95436	79115	08303	01041	20030	63754	08459	28364
55957	57243	83865	09911	19761	66535	40102	26646	60147	15702
46276	87453	44790	67122	45573	84358	21625	16999	13385	22782
55363	07449	34835	15290	76616	67191	12777	21861	68689	03263
69393	92785	49902	58447	42048	30378	87618	26933	40640	16281
13186	29431	88190	04588	38733	81290	89541	70290	40113	08243
17726	28652	56836	78351	47327	18518	92222	55201	27340	10493
36520	64465	05550	30157	82242	29520	69753	72602	23756	54935
81628	36100	39254	56835	37636	02421	98063	89641	64953	99337
84649	48968	75215	75498	49539	74240	03466	049292	36401	45525
63291	11618	12613	75055	43915	26488	41116	64531	56827	30825
70502	53225	03655	05915	37140	57051	48393	91322	25653	06543
06426	24771	59935	49801	11082	66762	94477	02494	88215	27191
20711	55609	29430	70165	45406	78484	31639	52009	18873	96927
41990	70538	77191	25860	55204	73417	83920	69468	74972	38712
72452	36618	76298	26678	89334	33938	95567	29380	75906	91807
37042	40318	57099	10528	09925	89773	41335	96244	29002	46453
53766	52875	15987	46962	67342	77592	57651	95508	80033	69828
90585	58955	53122	16025	84299	53310	67380	84249	25348	04332
32001	96293	37203	64516	51530	37069	40261	61374	05815	06714
62606	64324	46354	72157	67248	20135	49804	09226	64419	29457
10078	28073	85389	50324	14500	15562	64165	06125	71353	77669
91561	46145	24177	15294	10061	98124	75732	00815	83452	97355
13091	98112	53959	79607	52244	63303	10413	63839	74762	50289
73864	83014	72457	22682	03033	61714	88173	90835	00634	85169
66668	25467	48894	51043	02365	91726	09365	63167	95264	45643
84745	41042	29493	08136	09044	51926	43630	63470	76508	14194
48068	26805	94595	47907	13357	38412	33318	26098	82782	42851
54310	96175	97594	88616	42035	38093	36745	56702	40644	83514
14877	33095	10924	58013	61439	21882	42059	24177	58739	60170
78295	23179	02771	43646	59061	71411	05697	67194	30495	21157
67524	02865	38593	54278	04237	92441	26602	63835	38032	94770
58268	57219	68124	73455	83236	08710	04284	55005	84171	42596
97158	28672	50685	01181	24262	19427	52106	34308	73685	74246
04230	16831	69085	30802	65559	09205	71829	06489	85650	38707
94879	56606	30401	02602	57658	70091	54986	41394	60437	03195
71446	15232	66715	26385	91518	70566	02888	79941	39684	54315
32886	05644	79316	09819	00813	88407	17461	73925	53037	91904
62048	33711	25290	21526	02223	75947	66466	06232	10913	75336
84534	42351	21628	53669	81352	95152	08107	98814	72743	12849
84707	15885	84710	35866	06446	86311	32648	88141	73902	69981
19409	40868	64220	80861	13860	68493	52908	26374	63097	45052
57978	48015	25973	66777	45924	56144	24742	96702	88200	66162
57295	98298	11199	96510	75228	41600	47192	43267	35973	23152

参 考 文 献

[1] 贾俊平. 统计学[M]. 北京：中国人民大学出版社，2009.

[2] 莱文等. 商务统计学[M]. 北京：中国人民大学出版社，2006.

[3] 卡赛，贝耶. 统计推断[M]. 北京：机械工业出版社，2004.

[4] 李洁明，祁新娥. 统计学原理[M]. 上海：复旦大学出版社，2010.

[5] 安德森等. 商务与经济统计[M]. 北京：机械工业出版社，2010.

[6] 曾五一. 统计学[M]. 北京：中国金融出版社，2006.

[7] 汉拿根. 统计学[M]. 北京：经济管理出版社，2008.

[8] 凯勒，沃拉克. 统计学：在经济和管理中的应用[M]. 北京：中国人民大学出版社，2006.

[9] 郭志刚. 社会统计分析方法——SPSS 软件应用[M]. 北京：中国人民大学出版社，1999.

[10] 张梅琳. 应用统计学[M]. 上海：复旦大学出版社，2008.

[11] 何晓群. 现代统计分析方法与应用[M]. 北京：中国人民大学出版社，2007.

[12] 马立平，刘娟. 应用统计学[M]. 北京：首都经济贸易大学出版社，2011.

[13] 吴喜之. 统计学：从数据到结论[M]. 北京：中国人民大学出版社，2006.